COMMENT ON

'New Capabilities and Methods of the X12-ARIMA Seasonal-Adjustment Program."

Findley,Monsell,Bell,Otto and Chen. JBES,98

Agustín MARAVALL
Bank of Spain, 28014 Madrid, Spain

1. GENERAL REMARK

The program X-12-ARIMA contains significant improvements over X-11/X-11-ARIMA. Broadly, they can be classified into three groups—(1) the development of reg ARIMA, (2) some new options for the filters, and (3) more diagnostics. Clearly, (1) implies the belief that regARIMA models are appropriate for the series. Group (2) implies an increased filter flexibility, where the filter selection depends on the data. These improvements obviously represent a move toward a "model-based" (MB) approach, which is also implicit in the wording of the article [X-12-ARIMA "decomposes a . . . time series into a product of (estimates of) of a trend component (p. 129)"]. The distinction between a "theoretical" component and its estimator forms the basis of an MB approach. (Still, how is the theoretical component? What is the estimation criterion?) Perhaps the reasons for not moving all the way to an MB method are the old fears having to do with the need for experts and with computing time, and, of course, the power of inertia. I believe that the fears are not appropriate anymore because, for example, programs TRAMO and SEATS are fully model-based, fully automatic, and faster than X-12-ARIMA. In the comment, I will center on flexibility and diagnosis, and I will use as examples basic macroeconomic Spanish series.

2. FLEXIBILITY: IDEMPOTENCY AND SPURIOUS RESULTS

In the difficult field of finding "objective" criteria for comparison of seasonal-adjustment methods, there are two that seem unquestionably desirable. One is idempotency; that is, a seasonal-adjustment method applied to the seasonally adjusted (SA) series that it has produced should leave the SA series unchanged. The second requirement is that, when applied to white noise, the method should produce no spurious seasonality. Both properties show, of course, how flexible a filter is to adapt to the particular structure of a series.

With a reasonable automatic model identification (AMI) procedure, an MB approach would identify nonseasonal models in both cases. But let us move one step backward and compare X-12-ARIMA to the simplest MB procedure, whereby only a default model is considered—namely the so-called airline model,

$$\nabla\nabla_{12}x_t = (1+\theta_1 B)(1+\theta_{12}B^{12})a_t + \mu. \qquad (1)$$

Thus, X-12-ARIMA run by default is compared with a procedure that consists of fitting the default model and using this model to obtain the filters for the component estimates (equivalent to running SEATS by default).

For the Consumer Price Index, Figure 1 compares the seasonal component of the original series with those obtained for the SA series, using X-12-ARIMA both times. The seasonality in the SA series, although not large, is nevertheless disturbing. Figure 2 performs the same comparison for the MB procedure I consider (the default model

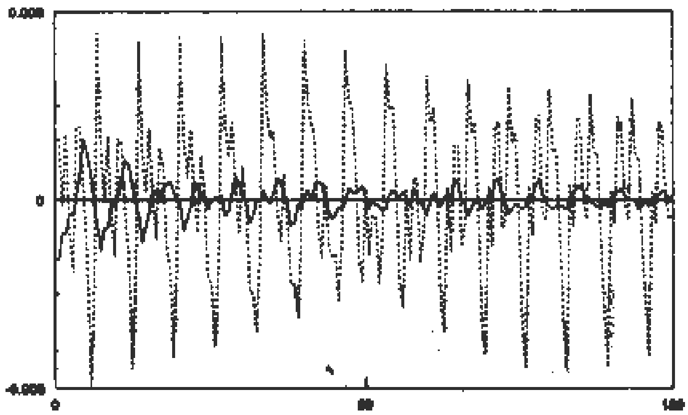


Figure 1. Idempotency: X-12-ARIMA: Seasonal Component: ——, Adjusted; - - -, Original.

© 1998 American Statistical Association
Journal of Business & Economic Statistics
April 1998, Vol. 16, No. 2

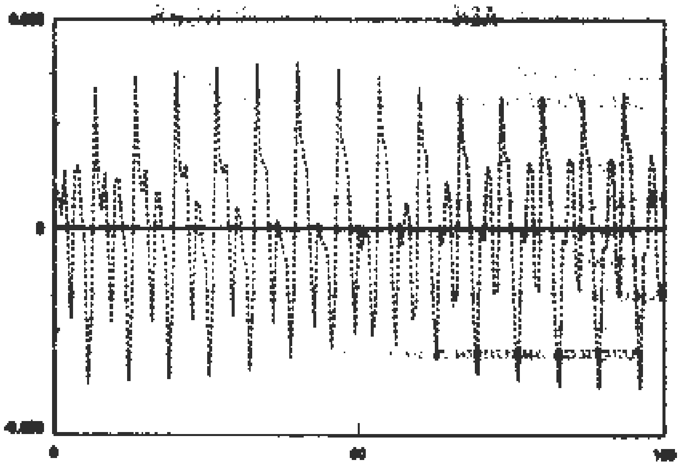


Figure 2. Idempotency: SEATS: Seasonal Component: ——, Original; - - -, Adjusted.

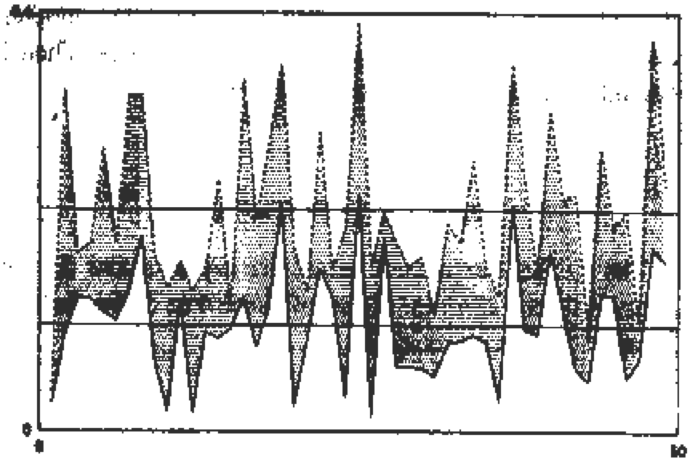


Figure 4. Spurious Seasonality: Variance [s(t)]: ——, var(SEATS); - - -, var(X-12-ARIMA).

yields a good fit). To all purposes, the estimated seasonal component in the SA series is now 0. Figure 3 compares the seasonality estimated in the SA series by both methods.

As for spurious seasonality, I generated 50 white-noise (0, 1) independent series and applied the two methods. Figure 4 compares the 50 variances of the estimated s_t series (the straight lines are the mean values: .211 for X-12-ARIMA and .101 for the MB method). For all 50 cases, the variance of the spurious seasonal component for X-12-ARIMA is considerably larger than that for the MB case.

The flexibility of the MB approach is explained by a simple feature. When model (1) is fit to the SA series, θ_{12} converges to -1 fast. Stopping its value at, say, $-.98$ (or $-.99$), the seasonal structure of the model in practice cancels out. Notice that, if deterministic seasonality were to be present, it would have been well captured, as shown in Figure 5, which magnifies the MB estimator of s_t. In a similar way, when Model (1) is fit to white noise, both θ_{12} and θ_1 tend to -1 so that the regular unit root in practice also disappears and the model reproduces white noise well. As a consequence, besides the good idempotency properties, in the MB case there is no need to worry about whether seasonality or trend are present; moreover, there is no need to worry about whether there may be deterministic or stochastic seasonality present: The model will handle it.

The default model is flexible enough to encompass a wide variety of simpler cases, but the robustness of the also extends to larger models. Figure 6 exhibits the seasonal factors obtained from application of the default model and of the model

$$\nabla^2\nabla_{12}x_t = (1+\theta_1 B + \theta_2 B^2)(1+\theta_{12}B^{12})a_t \qquad (2)$$

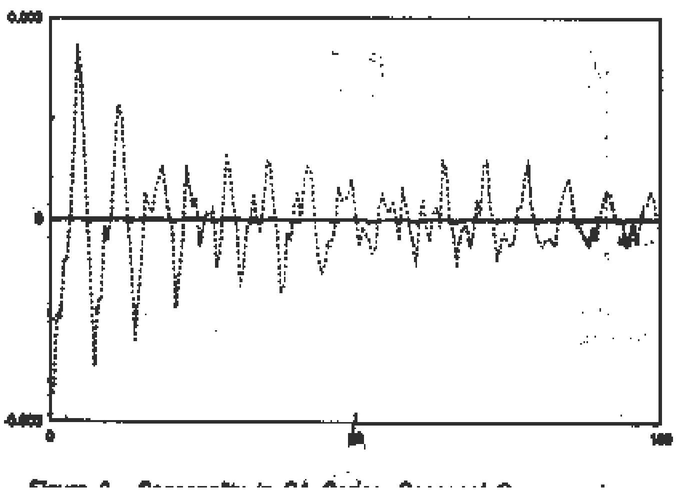
to the monetary aggregate liquid assets in the hands of the public. Because the default model provides an acceptable fit, Model (2) may be seen as the result of overdifferencing. The two sets of seasonal factors are indistinguishable and, again, this is because in estimation, one of the roots of the regular moving average (MA) polynomial goes to -1, and the effect of overdifferencing is cancelled out. Moderate overdifferencing causes, in practice, little damage.

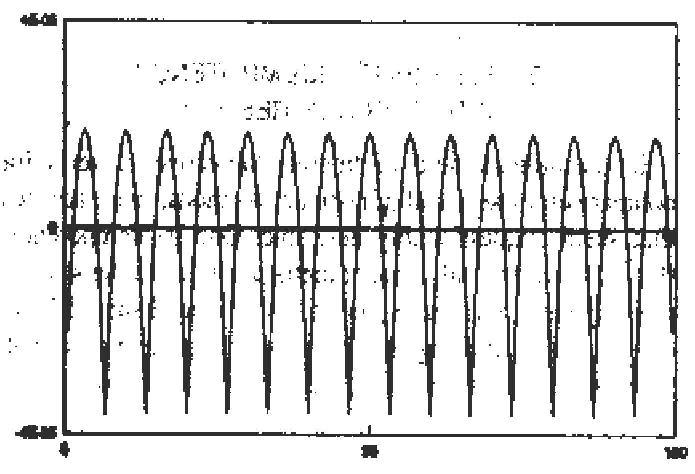
3. TESTING AND INTERPRETATION OF DIAGNOSTICS

The mixing in X-12-ARIMA of MB and non-MB features leads to some confusion. I shall illustrate this with some examples.

3.1 Testing for Trading Day and Easter Effects

The article states that the estimator of the irregular, "being an almost uncorrelated series" (p. 132), can be used for trading-day (TD) and Easter-effects (EE) estimation using simply ordinary least squares. Uncorrelatedness characterizes the component but certainly not the estimator, which can be strongly correlated (see Maravall 1995). Moreover, the estimator of the irregular (and of the SA series) is a non-





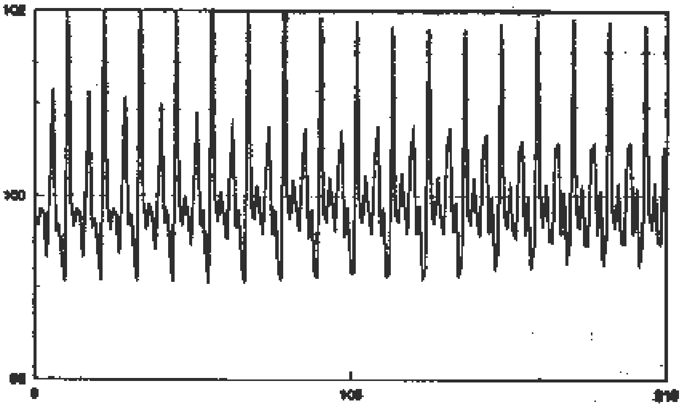


Figure 6. Seasonality and Overdifferencing: Seasonal Factors: ——, Overdifferencing; - - -, Default.

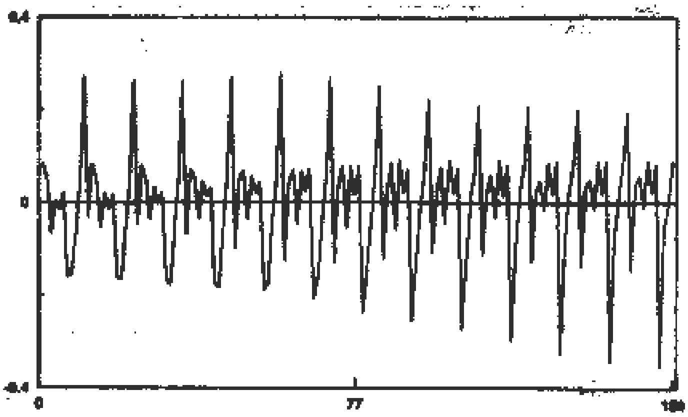


Figure 7. Seasonal Component: Exports.

invertible series and hence the use of finite autoregressive (AR) models to estimate its spectrum makes little sense. In this respect, thus, I agree with the authors; regARIMA seems the proper framework for TD and EE estimation (i.e., the MB way seems preferable to the old X-11 spirit).

3.2 Sliding-Spans and Revision-Histories Diagnostics

These are new diagnostics. The first one basically consists of running successively the program on overlapping subspans of the series, say $[x_{t-h}, \dots, x_t], [x_{t-h+12}, \dots, x_{t+12}], \dots, [x_{T-h}, \dots, x_T]$, and looking at the variation of $s_{t|t+j}$, the estimator of s_t obtained with the subspan finishing at $t+j$, as j increases. If that variation exceeds a limit k (recommended value of .03), month t is "unreliable"; if the percent of unreliable months is larger than 25%, then the series should not be adjusted (the variability is "much too high"). This type of diagnostics, in my view, tends to mix what should be the characteristics of a good extraction method with the analyst's wishful thinking concerning the properties of a series.

Consider the decomposition $x_t = s_t + n_t$, where s_t and n_t denote the (orthogonals) seasonal component and SA series, respectively. Assume that s_t is generated by a perfectly acceptable seasonal-component model of the type,

$$(1 + B + \cdots + B^{11})s_t = \theta_s(B)a_{st}, \qquad (3)$$

where a_{st} is white noise with variance V_s and the right side is a finite variance MA. Furthermore, assume that n_t follows some ARIMA model. Let us fix the models for s_t and n_t, except for V_s, which is systematically increased. The new seasonal components obtained are all in exactly the same way reasonable seasonal components. For V_s beyond some a priori fixed limit, why shouldn't the series be adjusted? (The diagnostic reminds me of the one whereby, if the average forecasting error is greater than 12%, the series should not be forecasted. If stock prices could be forecasted with 12% error, would it not be insane for investors not to forecast?) The variability that should be acceptable should be a function of the variability the series displays. This dependence is even implicit in the article when we are told that the recommended value of k will be too large for series with small seasonality and too small for series with large seasonal movements.

The authors recommend that the span length be at least as large as the length of the filter, and that one should look at the variability starting with concurrent estimation. Because the filter can then be completed in one direction, the variability in the successive estimators of s_t is, in essence, the total revision in the concurrent estimator. The MB approach permits us to address the variability issue in a more elegant and efficient manner (it also yields a more direct measure of the uncertainty in the measurement of s_t—namely, the standard error of the estimator.)

Let s_t^f and s_t^c denote the final estimator of s_t, obtained with the complete filter, and the concurrent estimator, $s_{t|t}$, respectively. The MB equivalent of the condition for not adjusting the series could be expressed as

$$\Pr[|s_t^f - s_t^c| > k] > .25 \qquad (4)$$

or, letting $d_t = s_t^f - s_t^c$ denote the revision in the concurrent

Table 1. Unreliability of Seasonal Component

	Exports (%)	Imports (%)
Probability that month t is unreliable	41	33
Frequency of unreliable months (sliding-spans diagnostic)	38	40

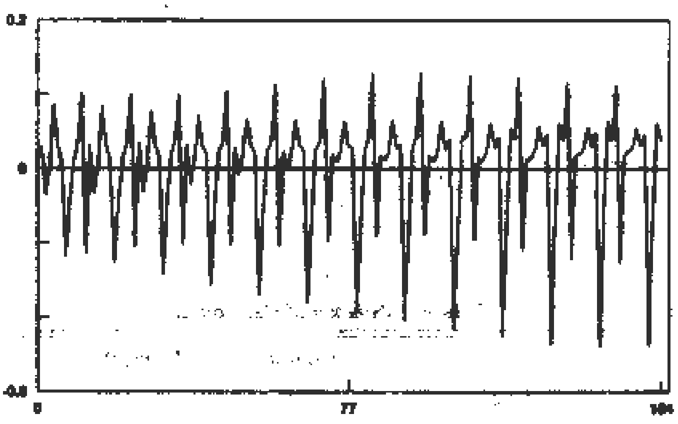


Figure 8. Seasonal Component: Imports.

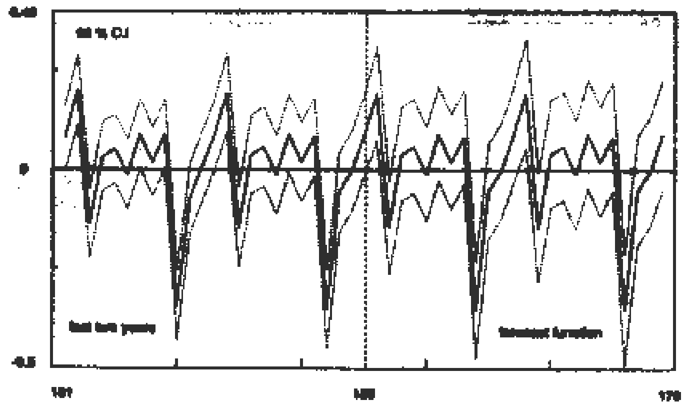


Figure 9. Seasonal Component: Exports.

estimator and assuming $k = .3$, $P(|d_t| > .03) > .25$. It is straightforward to express d_t as a linear combination of a_{t+j} with $j > 0$ (see Maravall 1995) from which the distribution of d_t can be easily obtained.

Using as examples the total exports and imports series, for which the default model (1) provides also good fits, the probability in the left side of (4) was computed, obtaining Table 1.

Thus X-12-ARIMA would conclude that neither of the two series should be adjusted. Figures 7 and 8 present the MB estimates of the two series of seasonal components. They display a "moving" structure but can be estimated nicely; moreover, as shown in Figure 9, they can also be forecasted well. In conclusion, the sliding-span exercise seems a questionable tool as a diagnostic on whether or not to adjust. In the MB approach, the sliding spans could be used indeed as a diagnostic: If the frequency of unreliable months is markedly different from the probability that a month is unreliable, then something wrong could be suspected. (For the exports and imports series, both computations are in clear agreement.)

A similar reflection applies to the revision-histories diagnostic. Given the stochastic structure of the series, there is an optimal revision process, and departures from it may be costly in terms of mean squared error. I agree with the authors that it may be preferable to use trends if revisions are smaller, yet the statement needs qualification. Two features of revisions are important, (1) the size of the revision and (2) the length of the revision period, and in general a smoother component does not decrease both. For the exports and imports series, Table 2 gives the variance of the revision in the concurrent estimator of the trend and of the SA series (the variance is standarized by setting $V_a = 1$). It is seen that the trend revision is larger in both cases. Table 3 shows the percent reduction in the variance of the revision after one year of additional data has become available.

Table 2. Variance of Revision Error

	Exports	*Imports*
Trend	.088	.093
SA series	.077	.061

Table 3. Reduction in Variance of Revision (%): One More Year of Data

	Exports	*Imports*
Trend	92	91
SA series	31	22

It is seen that the trend estimator converges much faster to the final estimator. This trade-off between size and duration is often found in practice, and it is difficult to say what is preferable, a moderate revision that takes many years to be completed or a large one that is removed in a few periods. Be that as it may, the imprecise criterion of having somewhat small revisions should be replaced by that of having optimal revisions (in the sense that minimum mean squared error estimators are obtained).

4. CONCLUDING REMARK

I congratulate the authors for their good work and hope they continue in the same direction (i.e., the MB direction). Then I think it is likely that X-13 will be an ARIMA MB method.

APPENDIX: MORE ON MISSPECIFICATION AND TESTING

In the author's reply to my comment (pp. 169–177) there are two points that deserve some further analysis.

A.1 Misspecification and Spurious Results

In my comment I mentioned that X-12-ARIMA would seasonally adjust white noise. In the reply, the authors point out that, in this case, the M and Q statistics would likely reject the decomposition. Fair enough, and their reply allows me to complete the argument.

I performed the following exercise. Twenty series of 120 observations were generated from model (1) with $\theta_1 = -.4$, $\theta_{12} = -.6$, and $a_t \sim$ niid $(0, \sigma_a = 1)$; as shown by Cleveland and Tiao (1976), for this model seasonal adjustment with X-11 is quite appropriate. X-12-ARIMA was applied to each x_t, with the series extended using the model chosen by regARIMA. Somewhat surprisingly, although the orders chosen for the seasonal polynomials were $(0,1,1)_{12}$ in all cases, for 12 of the series the orders of the regular polynomials were (2, 1, 2), for 6 they were (2, 1, 0), and only for 2 were the correct orders found. The automatic model-identification procedure of TRAMO [RSA = 3, which includes joint automatic outlier detection and correction and automatic model identification] selected always the $(0,1,1)(0,1,1)_{12}$ model. The difference may be a result of the bias toward overparameterization of the Akaike information criterion used by regARIMA; moreover, in my experience, the (2, 1, 2) specification is poorly behaved, unstable, and seldom needed.

Be that as it may, the decomposition obtained with X-12-ARIMA was for all 20 cases satisfactory and not a single M or Q statistic fell outside the acceptable range (the average Q value was .29). Denote by s_t(X12) the seasonal component produced by X-12-ARIMA. Twenty series were

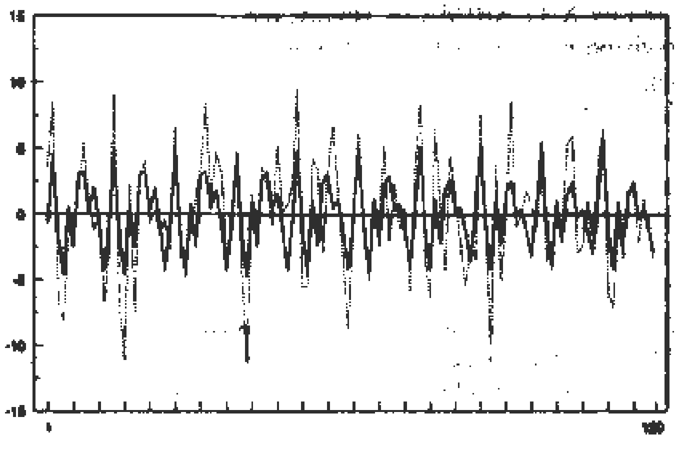


Figure A.1. Constructed Series: ——, X-12-ARIMA Seasonal Component; . . . , X-12-ARIMA Seasonal Component With Added Noise.

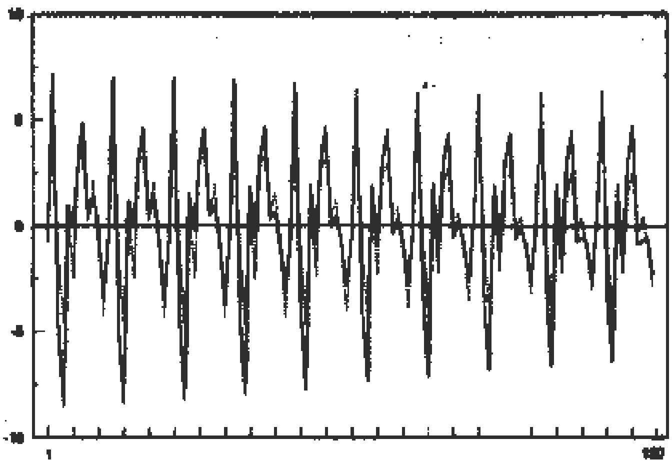


Figure A.2. Estimator of the Seasonal Component: X-12-ARIMA: ——, X-12 Estimator; . . . , Original Seasonal Component.

constructed by perturbing these seasonal components with noise, as in

$$z_t = s_t(\text{X12}) + n_t, \tag{A.1}$$

where n_t is niid with variance equal to 2 (the average variance for the 20 seasonal components was 5.4). Then, X-12-ARIMA was applied to the z series. For none of the series was a model chosen by regARIMA. More relevantly, the average number of unacceptable M statistics was 6 (range: 9–4), the average Q statistic was 1.56 (range: 1.71–1.34), and the decomposition was always rejected. As for sliding-spans diagnostics, for the cases in which it was computed (about half) the percentage of unstable months for the SA series varied between 94.8 and 100, and for its month-to-month changes, between 94.8 and 99. In conclusion, it is clear that if one uses X-12-ARIMA in the way "it was designed to be used" (and recommended by the authors), seasonal adjustment of the 20 z series has to be drastically rejected, and X-12 "can be regarded as the 'identity operator' with such series" (p. 173). Yet, by construction, the series consist of a dominant, perfectly sensible X-12-ARIMA seasonal component, which would be included untouched in the seasonally adjusted series. Figure A.1 presents one of the series; regarding our points of concern, all series displayed very similar features.

One may think that, although the variance of the seasonal component is nearly three times that of the noise, the amount of the latter is nevertheless too high and prevents estimation of the seasonal component. Using X-12 in "the way in which it was not designed to be used" (i.e., ignoring the diagnostics), let $\hat{s}_t(\text{X12})$ denote the seasonal component obtained for the z series. The variance of the estimation error $e_t(\text{X12}) = s_t(\text{X12}) - \hat{s}_t(\text{X12})$ is equal, on average, to .44 of the variance of $s_t(\text{x12})$. Besides, if $\hat{n}_t(\text{X12})$ denotes the SA series, the correlation between n_t and its estimator $\hat{n}_t(\text{X12})$ is, on average, .87 (range: .93–.82). All considered, $\hat{s}_t(\text{X12})$ may not be a great estimator, but it is certainly better than 0. The overwhelming rejection of the adjustment seems, for all 20 z series, undeserved. Figure A.2 contains the two series $s_t(\text{X12})$ and $\hat{s}_t(\text{X12})$ for the example considered.

As a consequence, if the outcome of X-12-ARIMA is left untouched, a spurious seasonal component would be removed from white noise; if the diagnostics are used (as recommended), no seasonality would be removed from (cuasi) purely seasonal series. In automatic use on many series, for the ones that are not close to a particular structure, spurious results are bound to appear with and without diagnostics.

Because the z series were constructed by adding noise to an X-12-ARIMA-produced seasonal component, it would seem that X-12-ARIMA should be more appropriate than an MB method to seasonally adjust the z series. This belief could be reinforced if one considers the stochastic structure of these constructed series. Given that the MB approximations to the X-11 seasonal component are all of the type (3), adding noise would yield (approximately) a model of the type

$$(1 + B + \cdots + B^{11})z_t = \theta_q a_t \tag{A.2}$$

—that is, an ARMA (11, q) model with $q \geq 11$, which is unlikely to be considered by any automatic model-identification procedure. Nevertheless, I tried TRAMO/SEATS on the z series, again in the automatic mode (RSA = 3). For 13 of the series, the model identified was of the type

$$\nabla_{12} z_t = (1 + \theta_{12} B^{12}) a_t; \tag{A.3}$$

for the rest of the series, it is given by Model (1), the airline model. The values of $\hat{\theta}_{12}$ vary between $-.69$ and -98. Even for the one with smallest absolute value, because $(1 - .69B^{12}) = (1 - .97B)(1 + .97B + \cdots + .97^{11}B^{11})$, the root $(1 - .97B)$ cancels in practice the root $(1 - B)$ in the factorization of $\nabla_{12} = (1 - B)(1 + B + \cdots + B^{11})$. Therefore, Model (A.3) simplifies into a model of the type (A.2), more in line with the generating model. For the airline-model cases, the maximum value for $\hat{\theta}_1$ was $-.95$ and, again, the MA root $(1 + \theta_1 B)$ cancels in practice the regular ∇, and the model simplifies into a model of the type (A.2). In fact, the in- and out-of-sample diagnostics obtained are in all cases acceptable. The 24-lag Ljung–Box statistics for residual autocorrelation had an average value of 23.5 (range: 35.1–14.8), and an out-of-sample forecast

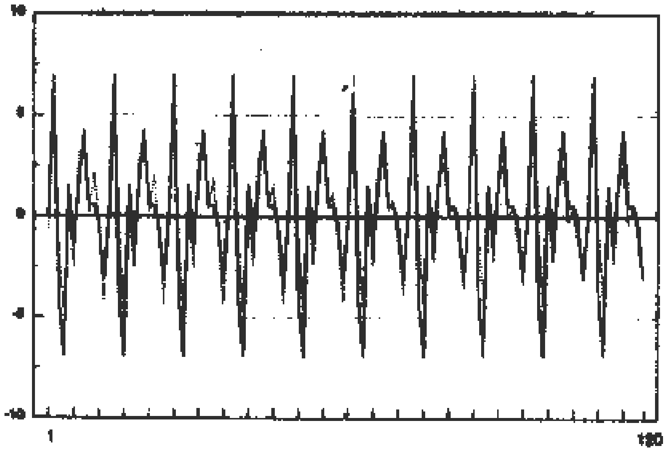


Figure A.3. Estimator of the Seasonal Component: TRAMO/SEATS: ——, TRAMO/SEATS Estimator; . . . , Original Seasonal Component.

F test (with 12 one-step-ahead forecast errors) equals, on average, .88 (range: 1.57–.37).

Letting $\hat{s}_t$(TS) and $\hat{n}_t$(TS) denote the seasonal and noise estimators obtained for the series z_t using the TRAMO/SEATS procedure, the variance of the error in the estimator of the seasonal (s_t(X12) − $\hat{s}_t$(TS)) becomes on average .23 (range: .32–.12), and hence the TS estimator cuts in half the estimation error of X-12-ARIMA. Moreover, the average correlation between the SA series n_t and its estimator $\hat{n}_t$(TS) jumps to .93 (range: .95–.87). In fact, *for all 20 cases,* the TS estimator of s_t(X12) was more precise than the one obtained with X-12-ARIMA, and the correlation between the adjusted series and its estimator was larger. Furthermore, as a result of having removed the noise better, the seasonal component obtained with TS was systematically more stable (see Fig. A.3). In summary, the flexibility of the model-based approach permits recovering an X-12-ARIMA-generated seasonal component that has simply been contaminated with some noise with considerably more accuracy than does X-12-ARIMA itself.

A.2 Sliding Spans and Revisions

In building the sliding-spans diagnostics, the authors recommend starting with the span that finishes with the last observation, choosing a span length close to the length of the filter, and considering a minimum of four spans (with starting dates of neighboring spans differing by a year). Consider the last observed period; the purpose of the sliding-spans diagnostic is to look at the variability of the factor for that period as new observations become available, starting with concurrent estimation, and, because most of the filter has been completed in four years, finishing with an estimator close to the final or historical one. The changes in the factor induced by new data are, by definition, the revisions. Of course, one can look at the full revision series or at the revisions produced every 12 months. Given that the latter is equivalent to systematic sampling of the former, the full series provides a more complete description of the variability. As for the summary statistics of that variability, one can use range statistics of the type considered by X-12-ARIMA or, simply, the standard error of the revision process. Naturally, the numerical values of different statistics will differ, although the general message they convey (whether the variability of the revisions is high or low) should be in agreement. This is indeed the case in the examples considered in my comment, as was shown in Table 1, and the analysis is relevant. In fact, looking at Table 1 of the author's reply (p. 174), it remains true that the X-12-ARIMA decomposition of the series would be rejected and X-12-ARIMA would once more act as an identity operator. Thus, my remark remains valid: The sliding-spans diagnostics can be misleading. As was the case with the seasonal-plus-noise series in Section A.1, if the choice is between adjusting with X-12-ARIMA or leaving the original series untouched, then the series should clearly be adjusted (here I also refer to "subject-matter experts"). The SA series is somewhat erratic, due to the presence of a large irregular component, but the seasonal and trend-cycle components are sensible and well behaved. Incidentally, Table 1 of the reply also shows that, despite the fact that the forecast extension of the series in X-12-ARIMA is made using the model provided by the automatic procedure of TRAMO, the sliding-spans statistics are systematically better for TS than for X-12-ARIMA. The difference is small because the two decompositions are relatively close, yet the systematic improvement reveals the larger stability of the TS components.

The danger of yielding, as we have shown, misleading results that characterizes the use as diagnostics of the sliding spans and M statistics in X-11-ARIMA–X-12-ARIMA is, in my view, a product of the "wishful thinking" component of the philosophy behind the diagnostics. This philosophy is implicitly based on the belief in how a decent series should be; unfortunately, many series are not decent. In the MB approach, this problem disappears. The equivalent of the sliding-spans diagnostics (in essence, to compare the theoretical and actual values of the unreliability probabilities) was shown in my comment to give the right answer (i.e., both values are close).

The criticism to the sliding-spans analysis only refers to its use in diagnostics (i.e., to adjust or not to adjust). In deciding, for example, among (nested) alternatives, I believe they can be of help, as the authors illustrate for the case of including or not a TD specification. Yet the example also illustrates the poor performance of the (differenced) SA series and irregular component spectra in detecting residual effects. Contrary to what the reply seems to say, automatic use of TRAMO with the two TD specifications closest to the one used for X-12-ARIMA (RSA = 6, 7) rejects the presence of a TD effect. In fact, I found it also surprising that, for 5 of the 20 z series generated in Section A.1, at least one of the two spectra indicated the presence of a TD effect.

ADDITIONAL REFERENCE

Maravall, A. (1995), "Unobserved Components in Economic Time Series," in *The Handbook of Applied Econometrics* (vol. 1), eds. H. Pesaran, P. Schmidt, and M. Wickens, Oxford, U.K.: Basil Blackwell, pp. 12–72.