Diagnostics for Model-Based Seasonal Adjustment

Roxanne M. Feldpausch, Catherine C. H. Hood and Kellie C. Wills
U.S. Census Bureau, Washington, D.C. 20233-6200

1. Background

Progress in seasonal adjustment depends on the development not only of methods that better account for the various components of time series, but also the development of better diagnostics. A successful seasonal adjustment can depend as much on the diagnostics as on the methods. In this paper, we try to identify promising diagnostics for model-based seasonal adjustment.

Currently there are a variety of diagnostics for model-based adjustments available including model fit diagnostics and stability diagnostics. In this paper we investigate which diagnostics are useful to determine the quality of the seasonal adjustment. We use simulated series to assess which diagnostics are associated with accuracy of an adjustment.

While most of this research should be applicable to a model-based adjustment performed by any software, we focus on SEATS adjustments. SEATS (Signal Extraction in ARIMA Time Series) is a program developed by Agustín Maravall and Victor Gómez to seasonally adjust time series using ARIMA model-based signal extraction techniques. SEATS uses signal extraction with filters derived from an ARIMA-type time series model that describes the behavior of the series. This method is based on work by Hillmer and Tiao (1982) and Burman (1980), among others. (Also, see Maravall (1993) and Gómez and Maravall (1997).)

The Census Bureau has a version of its seasonal adjustment package, X-12-ARIMA, which has access to the SEATS algorithms. The new program is temporarily called X-12-SEATS. The advantage of X-12-SEATS is that it has many of the diagnostics available in X-12-ARIMA along with all of the diagnostics available in SEATS. The main difference between SEATS and the SEATS module of X-12-SEATS is that SEATS uses conditional likelihood for estimating AR models and X-12-SEATS uses an exact maximum likelihood. X-12-SEATS is still under development and is not yet being released by the Census Bureau. For more information, see the paper by Monsell, Aston, and Koopman (2003). For this research we used X-12-SEATS.

1.1 Diagnostics

X-12-SEATS includes two stability diagnostics: sliding spans and revision history diagnostics. The sliding spans diagnostics were developed at the Census Bureau (see Findley, Monsell, Shulman, and Pugh (1990) and Findley, Monsell, Bell, Otto, and Chen (1998)). The purpose of the sliding spans diagnostics is to compare adjustments from overlapping subspans of the series.

The sliding span procedure looks at up to four spans of data. All the spans are the same length, where length is determined by the seasonal moving average parameter. The last span ends at the last point in the series. The second-to-last span starts (and ends) one year earlier than the last span and so on. If there are not enough data for four spans, the program will calculate three spans or two spans, when possible. X-12-SEATS calculates seasonal adjustments for each span of data separately, resulting in up to four different estimates of the seasonal adjustment for a large number of points. It looks at the maximum percent difference between the estimates. Maximum percent difference is calculated for the seasonally adjusted series, the month-to-month percent change in the seasonally adjusted series, and the seasonal factors.

The span length for a SEATS adjustment is determined by the seasonal moving average parameter. Span lengths are short when the estimated parameter is small and long when it is close to one. When the span length is too long, it can be impossible to obtain even two spans for comparison. If there is no seasonal moving average parameter, the span length is four years.

The revision history procedure computes a sequence of adjustments from truncated sets of data. An initial or concurrent adjustment is calculated at time t using data up to time t. A final adjustment is also calculated at time t using the full series of data. This allows the user to compare

revisions from the initial estimate to the most recent estimate. Revisions can be quantified by the mean and maximum absolute percent difference between the initial and final estimate for the seasonal adjustment and between the initial and final estimate of the month-to-month percent change for the seasonal adjustment.

X-12-SEATS estimates the sample autocorrelation function, sample partial autocorrelation function and the Ljung-Box Q statistics for lags one through 36 of the regARIMA model residuals. The Ljung-Box Q is a portmanteau test, that is, the Q statistic corresponding to the k^{th} autocorrelation tests whether the first k autocorrelations are zero, as white noise. At low lags, Q follows a chi-square distribution; however, at large lags it does not follow a chi-square distribution, so we ignore Q_{25}-Q_{36} in our analysis. See Ljung and Box, (1978) for more details.

X-12-SEATS also provides a test for normality of the regARIMA model residuals. The test for normality of the residuals has two parts: Geary's A and kurtosis. Newer versions of X-12-SEATS also include a test for skewness.

SEATS uses an "overestimation" and "underestimation" diagnostic for detecting residual seasonality (Maravall, 2003). Overestimation of a component indicates that too much variation has been assigned to that component. Underestimation of a component indicates that its estimate does not capture all of the variation. This diagnostic is calculated for the trend, irregular, seasonal factors, and the seasonally adjusted series. This diagnostic was shown to be negatively biased, toward indicating underestimation (Findley, Wills, Aston, Feldpausch, Hood, 2003).

Note that for model-based adjustments some diagnostics in X-12-ARIMA are not available. For a SEATS adjustment you cannot get the F-tests for stable and moving seasonality. Also, the M and Q Monitoring and Quality diagnostics developed by Statistics Canada are not available. There is no diagnostic comparable to M7 for determining whether or not a series is seasonal.

2 Methods

We used two types of simulated series to evaluate some of the diagnostics available in X-12-SEATS. We used a set of airline simulations and another set of simulations that were based on real-life series. We used these airline series to make sure that SEATS and the model-based diagnostics performed how we expected them to. We then used the simulations based on real-life series to determine how well the diagnostics performed.

2.1 Data used

We simulated 1,000 series based on the Box-Jenkins airline model. This model is of the form

$$(1-B)(1-B^s)Z_t=(1-\theta B)(1-\Theta B^s)\varepsilon_t$$

where Z_t is the logarithm of a seasonal time series, s is the number of observations per year ($s\geq 2$), B is the backshift operator, and ε_t is a random variable that is distributed as white noise. We simulated series with s=12, monthly time series. For each θ and Θ considered, with a fixed value for the variance of ε_t for all series, we obtained the ARIMA models for the trend, seasonal and irregular components produced by SEATS' canonical decomposition of the airline model. Then we simulated independent Gaussian series from each component's model (e.g. white noise with the canonical variance for the irregular component). The sum of series from the three component's simulations yields an airline model series with the prescribed parameters whose true seasonal decomposition components are known. The component series were exponentiated to achieve a multiplicative seasonal decomposition. For our simulations, we used values θ and Θ randomly chosen between 0 and -0.988.

We also simulated some series based on real-life series. We used 267 published series from the U.S. Census Bureau's Import/Export series to create 1400 simulated series. The Import/Export series represent a wide range of possible seasonal patterns. We used X-12-SEATS to generate trend/cycle, seasonal, and irregular components for the Import/Export series from both X-11 and SEATS adjustments. The trend/cycle, seasonal factors, and irregular from different series were then combined together to form a new series. We will refer to these series as the Import/Export simulations. For a more detailed explanation of the methods we used to simulate these series, see Hood, Ashley and Findley (2000). All simulated series had length 156.

2.2 Judging Accuracy

For simulated series it is possible to judge the accuracy of the seasonal adjustment based on how close X-12-SEATS comes to the true seasonal

adjustment. For the series simulated by combining known seasonal, trend, and irregular components, we know what the true seasonal adjustment should be. For these series, we can see how close X-12-SEATS comes to the truth by looking at the relative mean absolute deviation (RMAD)

$$RMAD = N^{-1}\sum_{t=1}^{N}\frac{|\, x_t - \hat{x}_t \,|}{x_t}$$

where N is the number of data points in the series, x_t is the actual seasonally adjusted series, and $\hat{x}_t$ is the estimated seasonally adjusted series.

3 Results

As we expected, X-12-SEATS did a very good job with the airline simulations. Table 1 shows the mean RMAD for both types of simulations. The RMAD was smaller for the airline series. The results for the airline data given below are for adjustments using the model determined by TRAMO. The results for the airline data with the airline model specified (and with X-12-SEATS-estimated parameters) and the airline data with the TRAMO model were very similar. The mean and standard deviation were the same when rounded to two decimal places. We expected the Import/Export simulations to be more like real-life series and that X-12-SEATS would have a more difficult time with these series.

Table 1 Average RMAD for Simulated Series

Simulation	Mean (Std. Dev)	Range
Airline	0.15 (0.05)	0.01 - 0.70
Import/Export	2.55 (2.33)	0.00 - 27.61

In the following sections, we discuss results for individual diagnostics.

3.1 Sliding Spans

To evaluate the stability diagnostics, we used only the Import/Export simulations. We only considered series that had seasonal peaks in the spectrum of the original series. For series of length 13 years, like ours, there will be four spans to compare only when $\Theta < 0.695$. We restricted our analysis to the 45.8% of the simulations whose Θ satisfies this inequality. Series with less than four spans were more likely to be inaccurate and pass sliding spans than series with four spans. This is to be expected, since there are fewer points to compare. For this analysis, we only considered series where sliding spans had four spans of data.

For a series to pass sliding spans, (a) the 75th percentile of the maximum percent difference for the seasonal factors should be less than three and (b) the 60th percentile of the maximum percent difference for the month-to-month percent change should be less than three.

For a subset of series, we found that the sliding spans diagnostic is strongly correlated with accuracy of the adjustment. However, sliding spans did not perform equally well for all types of models.

For series that do not have a seasonal model all the seasonal factors should be the same. These series with nonseasonal models should all pass sliding spans, since the maximum percent differences should all be zero. For these series, the sliding spans results are not informative since the seasonal factors should be the same for all the spans. We eliminated series with nonseasonal models from our analysis.

For series without a seasonal difference, sliding spans also did not perform well. Most of the time, the maximum percent differences were close to zero, similar to the nonseasonal models. The issues for series with models without a seasonal difference are similar to those of nonseasonal models, the seasonal adjustment does not change very much across the spans.

Sliding spans performed best when there was a seasonal difference in the model. Figure 1 shows the response of the seasonal factors versus RMAD for series having a model with a seasonal difference. The Pearson correlation between the RMAD and the sliding spans 75th percentile of the seasonal factors was 0.77. Figure 2 shows response of the month-to-month percent change in the seasonal factors versus RMAD for series having a model with a seasonal difference. The Pearson correlation between the RMAD and the sliding spans 60th percentile of the month-to-month percent change in the seasonal factors was 0.70.

Note that all of our simulations began in January and ended in December. This meant that every month was represented in the spans the same number of times.

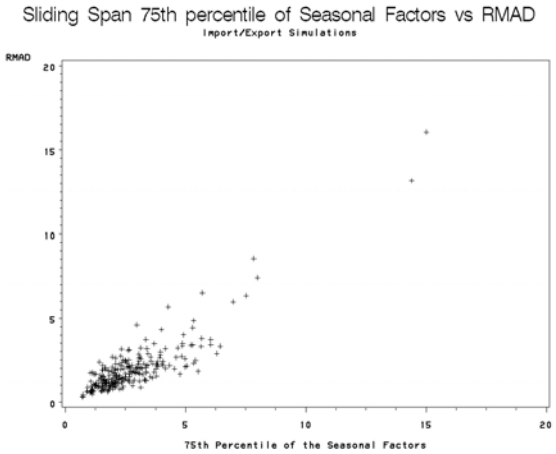

Figure 1 Response of the seasonal factor spans to accuracy

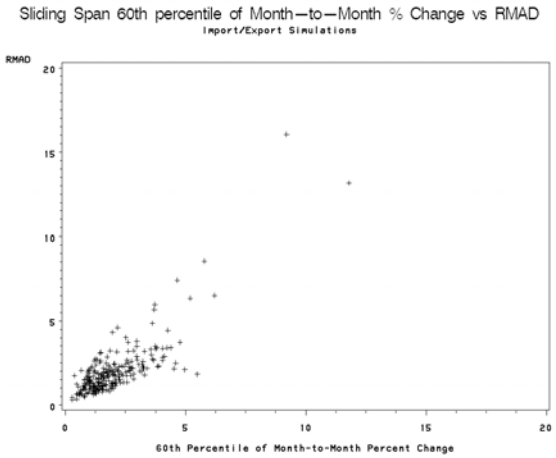

Figure 2 Response of the month-to-month percent change of the seasonal factor spans to accuracy

3.2 Revision History

The revision history successfully ran for 54.9% of the Import/Export simulations. Most of the time when revision history did not run, it was due to the model not converging at one or more points during the history run. The revision history diagnostics are also correlated with accuracy. Unlike sliding spans there is no cut-off value to indicate that a series passed or failed revisions history.

Similarly to sliding spans, revision history performed best for series that had models with a seasonal difference. Figure 3 shows the average absolute revisions of the seasonally adjusted series that had a seasonal difference in the model versus the RMAD. The correlation between the RMAD and the average absolute revisions of the seasonally adjusted series was 0.80. Figure 4 shows the average absolute revisions of the month-to-month percent change of the seasonally adjusted series with a seasonal difference in the model verse the RMAD. The correlation between the RMAD and the average absolute revisions of the month-to-month percent change of the seasonally adjusted series was 0.80.

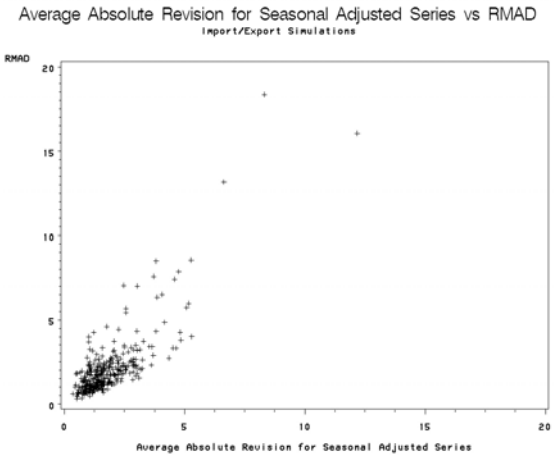

Figure 3 Response of the average absolute revisions of the seasonally adjusted series (revisions history) to accuracy

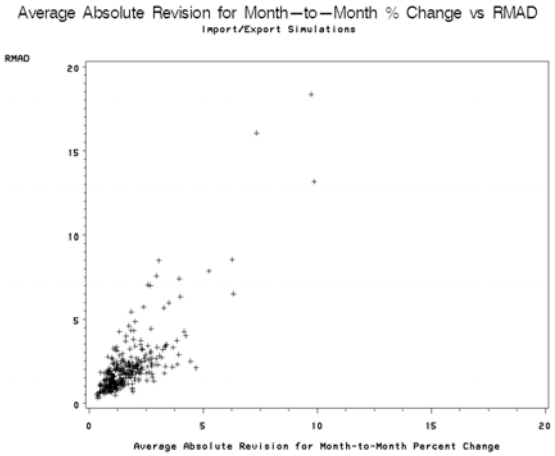

Figure 4 Response of the average absolute revisions of the month-to-month changes (revisions history) to accuracy

3.3 Ljung-Box Q

One of the model-based diagnostics that we examined was the Ljung-Box Q. To evaluate this diagnostic, we first used the simulated airline series with the airline model specified. We allowed X-12-SEATS to estimate the parameters. X-12-SEATS calculates Q_1 through Q_{36} for the regARIMA model residuals. These tests are each conducted at a 0.05 level. Normally in a situation where there are multiple tests, one would like to

take into consideration the issue of multiple comparisons. The Ljung-Box Q is complicated by the fact that the statistics are correlated. For example, if the ACF has a large peak at lag 16, this could cause Q_{16} through Q_{20} to fail. For this reason some people feel that it is best to look at only one Q. To this end, SEATS only computes Q_{24}.

We only considered Q_1-Q_{24}. We decided to concentrate on Q_{24} since this is the lag that is displayed in the SEATS output, and Q_{12} since 12 is our seasonal frequency. For lags less than or equal to 24, Table 2 shows the percentage of series where at least one Q failed, Q_{12} failed, and Q_{24} failed. Table 2 shows that the airline series failed Q_{12} and Q_{24} at approximately the expected 5% of the time. For the airline series, the median RMAD for those that failed Q_{12} (Q_{24}) and those that did not fail Q_{12} (Q_{24}) was the same, 0.15. Since the airline series produced the results we expected, we examined the Import/Export simulations. Table 2 also shows the results for the Import/Export simulations.

Table 2 Percentage of series where Q failed

Simulation	1+ Failing Q	Q_{12}	Q_{24}
Airline	30.8	5.8	5.3
Import/Export	75.0	11.5	14.7

Table 3 Median RMAD for Import/Export Simulations

	Q_{12}	Q_{24}
Failed	2.17	2.14
Did not fail	1.94	1.91

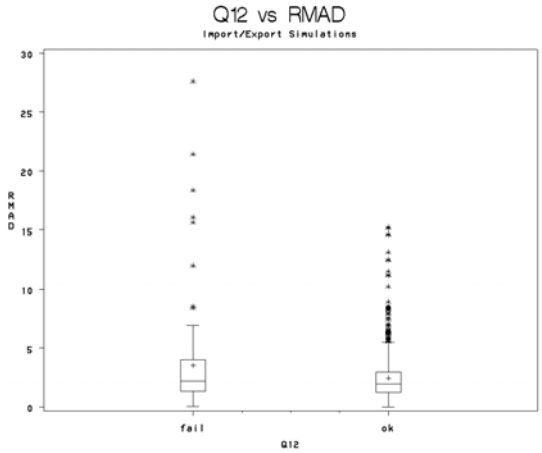

Figure 5 Box-plots of the RMAD for the series where Q_{12} failed and series where it did not fail.

Table 3 shows the median RMAD for those Import/Export simulations that failed Q_{12}. Figure 5 shows a box-plot of the RMAD for those series that failed Q_{12} and those that did not. The five series with the highest RMAD all failed Q_{12}. However, there were many series that did not fail Q_{12} and had high RMADs. Figure 6 shows a box-plot of the RMAD for those series that failed Q_{24} and those that did not.

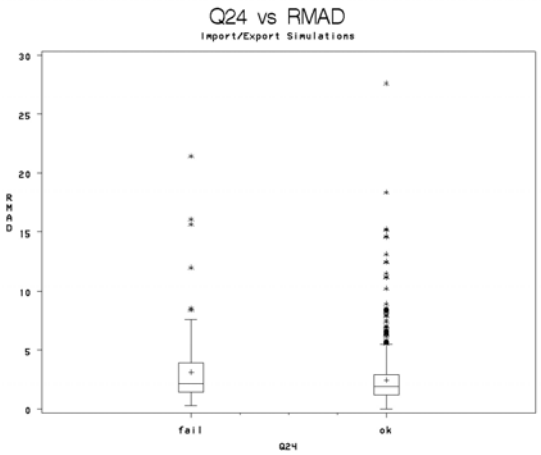

Figure 6 Box-plots of the RMAD for the series where Q_{24} failed and series where it did not fail.

3.4 Normality Test for Residuals

The test for normality of the RegARIMA model residuals was conducted at the 1% level. For the airline simulations, the test for normality of the residuals failed in 2.9% of the series. For the Import/Export simulations, the test for normality failed in 4.9% of series.

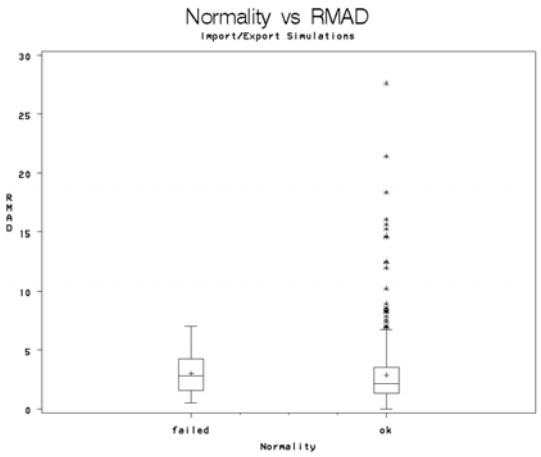

Figure 7 RMAD for series where the residuals failed the normality test and for series where the residuals did not fail the normality test.

The median RMAD for series that failed the normality test was 2.80. The median RMAD for series that did not fail the normality test was 2.14.

Figure 7 shows a box-plot of the RMAD for the series failing the normality test and for the series that did not fail the normality test. Based on the box-plot, it would be difficult to use the normality diagnostic to determine the accuracy of the adjustment. Both sets of simulations confirm that the failing the normality test is not a good indicator that the seasonal adjustment is inaccurate.

3.5 Overestimation/Underestimation Diagnostic

For the overestimation/underestimation diagnostics, we looked at the estimate minus the estimator. A negative value indicates underestimation; a positive value indicates overestimation. We would expect negative values for this diagnostic to be negatively correlated with the RMAD and positive values to be positively correlated with the RMAD. We examined this diagnostic using the airline simulations.

Figure 8 shows the overestimation/ underestimation diagnostic for the trend versus the RMAD. We looked at the correlation between negative values of the trend overestimation/underestimation diagnostic and the RMAD and the positive values and RMAD separately. There was no correlation between either the negative or positive values of the trend overestimation/underestimation diagnostic and the RMAD. For the series where this diagnostic was calculated, only 27.1% of the series had positive values.

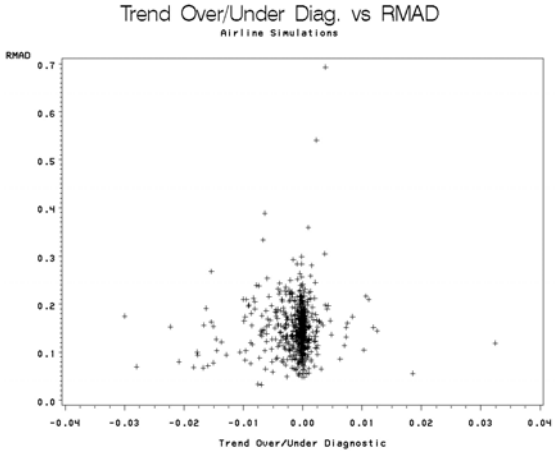

Figure 8 Trend Overestimation/Underestimation Diagnostic versus RMAD

Figure 9 is a graph of the irregular overestimation/underestimation diagnostic versus the RMAD. This diagnostic should be centered around zero. The graph clearly shows that this diagnostic is biased. Only 0.9% of the series had a positive value for the irregular overestimation/underestimation. There was some correlation between the negative values of this diagnostic and the RMAD; however, it was in the wrong direction. The correlation between negative values of the irregular overestimation/ underestimation diagnostic and the RMAD was 0.38.

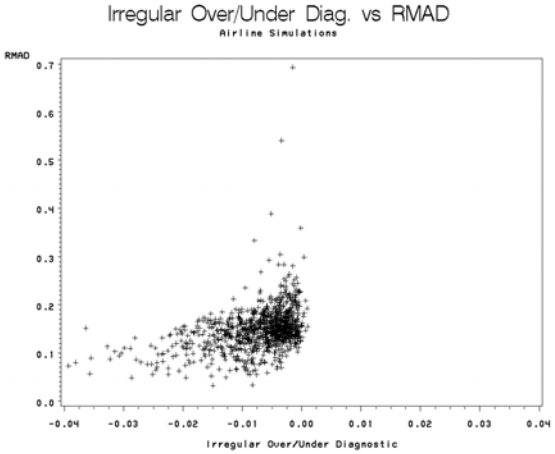

Figure 9 Irregular Overestimation/Underestimation Diagnostic versus RMAD

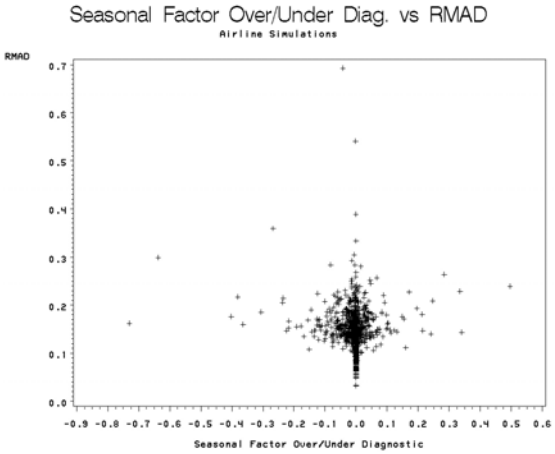

Figure 10 Seasonal Factor Overestimation/ Underestimation versus RMAD (note that this graph has a different scale than the other overestimation/underestimation graphs)

Figure 10 is a graph of the seasonal factor overestimation/underestimation diagnostic versus the RMAD. There was some correlation between the negative values of this diagnostic and the RMAD. There was also correlation between the positive values of this diagnostic and the RMAD. The correlation between negative values of the seasonal factor overestimation/underestimation diagnostic and the RMAD was –0.21; the

correlation between the positive values and the RMAD was 0.26. For series where this diagnostic was calculated, 29.1% of the series had positive values.

We decided to investigate the seasonal factor overestimation/underestimation diagnostic a little more using the Import/Export simulations. There was no correlation between the negative values of this diagnostic and the RMAD. There was a small correlation between the positive values of this diagnostic and the RMAD, 0.16 (p-value=0.038). For series where this diagnostic was computed, 24.5% of the series had a positive value. In general, the overestimation/ underestimation diagnostics did not do a good job indicating which series had an accurate adjustment.

4 Conclusions

The stability diagnostics are better indicators of accuracy of the adjustment than the model fit diagnostics. We recommend using the stability diagnostics to determine whether or not an adjustment is of acceptable quality. However, a problem with the stability diagnostics is that for some series it is impossible to obtain the stability diagnostics. For sliding spans, there may not be enough data to obtain the minimum of two spans. For revision history, the model may not converge at all of the points.

This paper focuses on using diagnostics to determine the quality of an adjustment. In the future, we would like to investigate which diagnostics are useful to look at when you are comparing two different adjustments. We also intend to examine the unbiased modifications of the overestimation/underestimation presented in Findley, McElroy, Wills (2004).

5 References

Burman, J.P. (1980), "Seasonal Adjustment by Signal Extraction," *Journal of the Royal Statistical Society*, Ser. A, 143: 321-337.

Findley, D.F., B.C. Monsell, H.B. Shulman, and M.G. Pugh (1990), "Sliding Spans Diagnostics for Seasonal and Related Adjustments," *Journal of the American Statistical Association*, 85:345-355.

Findley, D.F., B.C. Monsell, W.R. Bell, M.C. Otto and B.C. Chen (1998), "New Capabilities and Methods of the X-12-ARIMA Seasonal Adjustment Program" (with discussion), *Journal of Business and Economic Statistics*, 16: 127-176.

Findley D.F., K.C. Wills, J.A.D. Aston, R.M. Feldpausch, and C.C. Hood (2003), "Diagnostics for ARIMA-Model-Based Seasonal Adjustment," *2003 Proceedings of the American Statistical Association,* Business & Economic Statistics Section [CD-ROM], Alexandria, VA: American Statistical Association.

Findley, D.F., T.S. McElroy, and K.C. Wills (2004), "Modifications of SEATS' Diagnostic for Detecting Over- or Underestimation of Seasonal Adjustment Decomposition Components," to appear in *2004 Proceedings of the American Statistical Association,* Business & Economic Statistics Section [CD-ROM], Alexandria, VA: American Statistical Association.

Gomez, V. and A. Maravall (1997), "Program TRAMO and SEATS: Instructions for the User, Beta Version," Banco de Espana.

Hillmer, S.C. and G.C. Tiao (1982), "An ARIMA-Model-Based Approach to Seasonal Adjustment," *Journal of the American Statistical Association*, **77,** 377, pp. 63-70.

Hood, C.C, J.D. Ashley, and D.F. Findley (2000). "An Empirical Evaluation of the Performance of TRAMO/SEATS on Simulated Series." *Proceedings of the American Statistical Association, Business and Economic Statistics Section,* American Statistical Association: Alexandria, VA 171-176.

Ljung, G.M. and G.E.P. Box (1978) "On a measure of lack of fit in time series models", *Biometrika*, **65**, 2, pp. 297-303.

Maravall, A. (1993), "Unobserved Components in Economic Time Series," *Handbook of Applied Econometrics*, ed. M.H. Pesaran, T. Schmidt, M. Wickens, Oxford Basil Blackwell.

Maravall, A. (2003), "A Class of Diagnostics in the ARIMA-Model-Based Decomposition of a Time Series," Memorandum, Bank of Spain.

Monsell, B.C., J.A..D. Aston and S.J. Koopman (2003), "Toward X-13?**"** 2003 Proceedings of the American Statistical Association, Business & Economic Statistics Section [CD-ROM], Alexandria, VA: American Statistical Association.