Outlier Selection for RegARIMA Models

Kathleen M. McDonald-Johnson and Catherine C. Hood
U.S. Census Bureau, ESMPD Time Series Methods Staff, Washington, DC 20233-6250

KEY WORDS: seasonal adjustment; regression model with ARIMA errors

1. Abstract

In most data applications, statisticians must identify and estimate outlier effects. When doing seasonal adjustment, we are concerned that outliers may interfere with estimation of seasonal effects. By removing outlier effects, we hope to produce the best possible seasonal adjustment. The autocorrelation structure of time series differs from that of other types of data, so the outlier selection techniques also must be different. Using a large sample of economic time series from the U.S. Census Bureau, we fit regARIMA models (regression models with ARIMA errors) to the data with the X-12-ARIMA seasonal adjustment program. Our research simulated production as we added data one month at a time, refitting the regARIMA models for each run. We looked at the performance of automatic outlier identification when we raised or lowered the critical value, and we compared that to visual outlier selection methods. We expected our visual selection methods to improve on automatic outlier identification, but we concluded that the current automatic identification procedure was generally the best method.

After presenting background material, we describe our outlier identification methods, explain the regARIMA model diagnostics for the method comparison, and give our results.

2. Background

2.1 X-12-ARIMA

The Census Bureau uses a program called X-12-ARIMA to produce seasonally adjusted data. X-12-ARIMA is the most recent seasonal adjustment program in the X-11 line (Findley, Monsell, Bell, Otto, and Chen 1998). It follows X-11, developed at the U.S. Census Bureau, and X-11-ARIMA and X-11-ARIMA/88, developed at Statistics Canada.

One major improvement of X-12-ARIMA over X-11 is the use of regARIMA models—regression models with ARIMA errors—to estimate calendar effects or outlier effects with predefined or user-defined regressors. X-12-ARIMA uses regARIMA models to preadjust a series before seasonal adjustment by removing effects such as trading day, moving holidays, and outliers.

X-12-ARIMA has four predefined outlier types that users can specify as regression variables: a) additive outlier (AO), also called a point outlier, appropriate when one data point is unusual; b) level shift (LS), used when the level of the series changes suddenly; c) temporary change (TC), like an AO except that the effect is strong in one month and fades away exponentially in following months; and d) ramp, used when the level of the series changes more gradually—like an LS effect spread over a few months. Although users can specify ramps in the regression, the X-12-ARIMA automatic outlier identification procedure can identify only AOs, LSs, and TCs, so we used only these three types in this research.

X-12-ARIMA's automatic outlier identification procedure selects outliers by comparing regressor t-values to a critical value. The user can set a critical value or use the program's default value. Under the default settings, X-12-ARIMA adds one outlier at a time to the model. For each month of data, X-12-ARIMA calculates t-values for each type of outlier that the user has asked for (in this case, AO, LS, and TC). If at least one t-value has an absolute value greater than the critical value, the program adds the regressor (by type and date) for the outlier with the maximum absolute t-value, then recalculates the model estimates and outlier t-values. The program repeats the process of comparing absolute t-values to the critical value and adds outlier regressors one at a time, after each estimation of the model, until it finds no more absolute t-values that are greater than the critical value. During the identification phase, the program uses a robust estimate of the variance. When it cannot identify any additional outliers, it starts a backward deletion phase. After estimating the model including all identified outliers in the model, if at least one t-value of the model's outlier regressors has an absolute value less than the critical value, X-12-ARIMA removes the outlier regressor with the least significant t-value and estimates the model again. The program continues to remove outliers one at a time as long as at least one has an absolute t-value less than the critical value. During backward deletion, X-12-ARIMA uses the usual (nonrobust) residual variance estimate. The difference in variance affects the t-values, making the final values different from those calculated during the identification phase.

2.2 Strategies for Outlier Selection

One of the best strategies for selecting outliers in time series is to identify sources of outliers, such as strikes or severe weather. Such events are reason to add an outlier regressor to the regARIMA model, at least provisionally. However, without sufficient knowledge of the series it may not be possible for this type of intervention, so the automatic outlier procedure becomes very useful. With this procedure though, there are several different strategies for setting the critical value.

X-12-ARIMA sets the default critical value automatically according to the length of the series. This is a reasonable approach because the comparisons to the

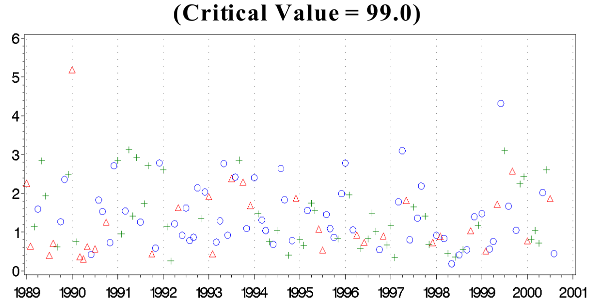
---

critical value involve several t tests, and the significance depends on the number of data points (Ljung 1993).

Another choice of critical value involves a threshold analysis. Using an automatic outlier identification procedure, the user completes several runs, each time lowering the critical value by a small amount (say 0.1), until the procedure identifies too many outliers (as determined by the user). The final critical value is the threshold just greater than the value that produced too many outliers. In the past we have seen examples where lowering the critical value this way improved the modeling diagnostics. In practice, it takes too much time because the user must run the program many times for each series.

In typical regression analysis, with independent data (that is, not time series data), plotting residuals from the regression can help identify outliers. Although we use regression with time series data, we cannot use the same method. Because X-12-ARIMA computes t-values for each outlier type for every month we can use those t-values to identify likely outliers. We plotted the maximum absolute t-value for each month.

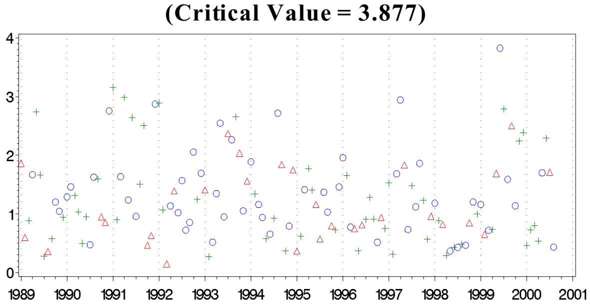
Figure 1. Maximum Absolute T-value (Critical Value = 99.0)



Grid Lines at January
Circle is AO, Plus Sign is LS, Triangle is TC

Figure 1 shows the absolute t-values that resulted from an X-12-ARIMA run in which we set the critical value very high so we would not identify any outliers. The values range from very close to zero to greater than five.

Figure 2. Maximum Absolute T-value (Critical Value = 3.877)



Grid Lines at January
Circle is AO, Plus Sign is LS, Triangle is TC

Figure 2 shows results for the same series after running X-12-ARIMA with the default critical value (3.877). Because X-12-ARIMA assigns a t-value of zero to identified outliers, generally they will not appear in the graph. In this case X-12-ARIMA identified January 1990 as a TC and set the t-value to zero, so the AO t-value is the maximum for January 1990. After looking at Figure 2, a user may decide to add an AO regressor at June 1999, as that t-value stands out from the rest of the graph.

3. Methods

3.1 All Methods

For this study we chose 206 United States Import/Export monthly time series that the Census Bureau's Foreign Trade Division (FTD) adjusts for seasonality. For some series, a simple model fits well. Others require complicated models, and outliers are often an important part of the model. Outliers can be caused by outside influences (such as a sudden rise in oil prices) or may result from definition changes as required by the Bureau of Economic Analysis.

Once a year, FTD staff members carefully model the series. We used their regARIMA models, excluding their outlier choices, as we identified outliers independently. Because we used predetermined regARIMA models, our resulting outlier choices were biased.

We looked only at individual series diagnostics, not at diagnostics for the resulting aggregate series, such as Total Imports, Total Exports, and Balance of Trade.

Our initial run included data from January 1989 to August 2000. We performed subsequent runs adding one month of data each time. These subsequent runs included revised data for the previous endpoint plus one month of preliminary data with the exception of the September 2000 data set. Unfortunately, we did not have original preliminary September data, so that data set included revised September values. Our final run included data from January 1989 to January 2001. We wanted to see how well the different outlier identification methods performed under conditions simulating production.

We devised outlier selection methods that we designated by letter, A - D. We also included a method with no outliers (Method Z). For all the methods, we set the initially-identified outliers to be regression variables for all subsequent runs. For subsequent runs we added a month of data and then ran X-12-ARIMA including the initial outliers as part of the model. We did not change the regARIMA model settings between runs. In addition to setting initial outliers, we performed automatic outlier identification on the full data span, although critical values differed from method to method.

3.2 Method Z: Zero Outliers

Method Z was not an outlier identification method, but for purposes of comparison we ran X-12-ARIMA with the automatic outlier identification procedure but a very high critical value (99.0). We wanted the results of outlier identification but no outliers. We used the results of Method Z for other methods. We graphed the t-values from the initial run of Method Z for Methods A and C.

3.3 Method A: Visual Outliers

From the maximum absolute outlier t-value graphs resulting from Method Z, we subjectively identified visual outliers without regard to the magnitude of the t-values. After adding regressors to the model for those visual outliers, we ran X-12-ARIMA for the initial data set and then for subsequent data sets.

This method became complicated when we put it into practice. For example, some series had neighboring outliers that made selection more difficult. Our solution was quite naive: when selected TCs or LSs occurred only one month away from another identified outlier of any type, we selected the outlier with the greater absolute t-value. We selected both outliers if they occurred two or more months apart. For series with several neighboring outliers including TCs and LSs, we selected the outlier with the greatest absolute t-value. However, we included all selected neighboring outliers if they were all AOs. Our predetermined outlier set (that we set to be regressors) resulted from applying this neighboring-outlier principle to the earlier selections.

The actual outlier set for each run consisted of our predetermined outliers and outliers that X-12-ARIMA identified automatically. We allowed automatic outlier identification because we were concerned that identifying some outliers would affect the significance of other t-values that were not obvious in the graphs. We did not set any of the automatically identified outliers to be regressors. The critical value for automatic identification was the maximum of the default critical value and the visual critical value (described under Method C).

3.4 Method B: Automatic Outlier Identification

We ran X-12-ARIMA for all series using the automatic outlier identification procedure with the default critical value. The only difference from the default settings was that we identified TCs as well as AOs and LSs. We added regressors for the identified outliers and ran X-12-ARIMA for the subsequent data sets. We used the default critical value for automatic identification.

3.5 Method C: Visual Critical Value

From the same graphs we used for Method A (resulting from Method Z), we set a visual critical value, a subjective choice of the value at which t-values begin to look like outliers when compared to other values in the graph. We set the value without regard to magnitude, with the constraint that the new critical value must be 3.0 or more. We ran X-12-ARIMA for the initial data set for each series using the new visual critical value, added regressors for the initially identified outliers, then ran X-12-ARIMA for subsequent data sets. We used the new visual critical value for automatic identification.

We allowed only one decimal place for our critical values, so none equaled the default critical values. Six times we chose 3.9 as the visual critical value, essentially equal to the default value for the initial run (3.877).

3.6 Method D: Visual Outliers After Automatic Identification

Method D was a combination of Methods A and B. X-12-ARIMA identified outliers for 91 series in the initial run for Method B. For those series we graphed the resulting maximum absolute t-values. (Because X-12-ARIMA assigns a t-value of zero for every *identified* outlier, the resulting graphs don't have extremely large t-values.) We identified visual outliers and set those to be regressors (in addition to the outliers previously identified). For the series with visual outliers we ran X-12-ARIMA again for the initial data set and then for subsequent data sets. We used the default critical value for automatic identification.

Of the 91 series with outliers in Method B's initial run, we identified visual outliers in only eight series; the other 83 duplicated Method B. For the remaining 115 series, Method D duplicated Method A.

3.7 Program Information

All of our regARIMA model estimates are from X-12-ARIMA version 0.2.8 for PC, dated May 16, 2001.

We used several SAS® programs to complete this research. Most prominently we used X-12-Graph (Hood 2001), a companion program to X-12-ARIMA that is written in SAS. X-12-Graph produced all of our t-value graphs and forecast error history graphs. We wrote additional SAS programs to automate and organize outlier identification. One program inserted our visual outliers, or for Methods B, C, and D the initially identified outliers, into the X-12-ARIMA input specification files as regression variables. The same program ensured that we used the appropriate critical value for automatic identification. We stored our selections in SAS data sets for easy comparisons.

4. Diagnostics to Determine Best Method

4.1 Model adequacy

To determine regARIMA model adequacy, we used familiar model adequacy diagnostics: spectra and Ljung-Box Q statistics.

Spectrum diagnostics indicate the presence of seasonal or trading day effects in a series. For monthly series, seasonal frequencies occur at k/12 cycles/month for $1 \le k \le 5$. We can detect trading day effects at frequencies 0.348 and 0.432 cycles/month (Cleveland and Devlin 1980). Spectrum graphs mark these frequencies for easy detection of the effects. A significant peak in the spectrum graph of the regARIMA model residuals at any of the seasonal or trading day frequencies is a signal of possible residual seasonality or trading day effects and can signify a lack of model adequacy. We quantify significance in units called stars. A peak of six or more stars is considered visually significant (Findley et al. 1998). X-12-ARIMA produced our estimates of the spectral peak sizes.

For us, failure of the residual seasonality test was a peak of six stars or more at k/12 cycles/month for $1 \le k \le 4$. We limited failure to those frequencies because 5/12 does not occur at a natural division of the year like the other seasonal frequencies. We did not want series to fail the diagnostic based solely on that frequency.

We also looked for trading day peaks. Because frequency 0.432 is considered appropriate for inventory series, we chose not to use it with our series. We looked for peaks only at frequency 0.348. Staff members in FTD have spent much time determining whether or not each series needs a trading day adjustment. Looking for trading day peaks was only to see which outlier set better supported the FTD decision.

In addition to spectrum results, we used the Ljung-Box Q diagnostic. Ljung-Box Q statistics are Chi-square lack-of-fit statistics for model residuals based on the autocorrelation function. Lags with a p-value less than 0.05 can indicate a lack of fit for the model. The most important lags are seasonal lags (12, 24, 36, etc. for monthly series) (Ljung and Box 1978). For us, failure was either a) a significant Q statistic at lag 12 or b) ten or more significant Q statistics for lags 1 through 24.

4.2 Outlier T-values and Additional Outliers

We expected the best outlier identification method to produce outliers that continued to be significant over time. The critical value was our significance measure.

The critical values corresponded to the identification method. For Methods B and C we had a critical value that we had used for outlier identification. For Method B that was the default critical value, and for Method C it was the visual critical value. The X-12-ARIMA default critical value changes with the addition of data points, so the measure for Method B changed with each added month of data. Methods A and D were somewhat different because we set visual outliers to be regression variables, not using a critical value for all identification. A logical measure for Method A was the visual critical value because we had chosen that value from the same graphs that we used to choose outliers for Method A. Method D did not have a definite critical value because we used the default critical value to identify most of the initial outliers, but t-values for the visual outliers were likely less than that critical value. (Exceptions could occur because of the different variance used in the identification and backward deletion steps.) Our significance measure for Method D was the minimum of the visual critical value and the default critical value.

In addition to checking significance of the outlier t-values over time, we looked at how sufficient the identification methods were. In production work, we normally use automatic outlier identification only for data that we add to the end of the time series. In our research, however, we used the automatic procedure to check how often X-12-ARIMA identified additional outliers within the *initial* model span. We would expect a good identification method not to select many additional outliers in the span that has already undergone outlier identification.

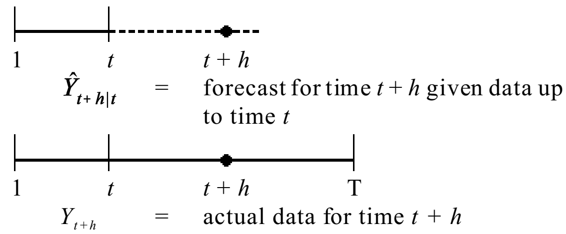
4.3 Forecast Performance

Some users consider forecast performance to be the most important criterion when choosing between competing regARIMA models. Forecast performance is important because in the current context of X-12-ARIMA, the regARIMA model adjusts series for regression effects like trading day and outliers but also provides forecasts for use with the symmetric X-11 seasonal moving averages.

X-12-ARIMA can compute a forecast error history from the regARIMA model estimation on a sequence of runs from truncated data sets (Findley et al. 1998). We can compare forecast errors of different models over time.

Let Y_n be a time series defined for n = 1, 2, ... ,T. $\hat{Y}_{t+h|t}$ is the h-step-ahead forecast of Y calculated using Y_1, Y_2, ... , Y_t, where $t \leq$ T. Because we use only the data up to time t to estimate the model coefficients and to calculate the forecast, it is an out-of-sample forecast. However, we also know the true value for Y_{t+h}, for $t \leq \mathrm{T} - h$.

Figure 3. Out-of-sample Forecast Error Illustration



Because we know the true value, we can compute the forecast error, $Y_{t+h} - \hat{Y}_{t+h|t}$, for $t_0 \leq t \leq \mathrm{T} - h$, where t_0 is the initial truncation point.

X-12-ARIMA can calculate forecasts for many overlapping time series, starting with the series that begins at the first point of data (time 1) and ends at some point t_0 and continuing through the series that begins at time 1 and ends at T − h. Each subsequent time series has one additional point of data at the endpoint (t_0 + 1, t_0 + 2,..., T − h). X-12-ARIMA estimates the model parameters for each time series of time 1 to time t, $t_0 \leq t \leq \mathrm{T} - h$.

To compare two models, let $\hat{Y}^{(i)}_{t+h|t}$ be model i's h-step-ahead forecast of Y_{t+h} at time t, i=1, 2. The squared errors are $(Y_{t+h} - \hat{Y}^{(1)}_{t+h|t})^2$ for Model 1 and $(Y_{t+h} - \hat{Y}^{(2)}_{t+h|t})^2$ for Model 2.

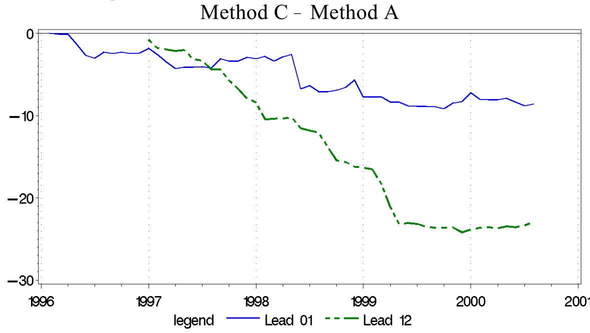
For given h and t_0, we can plot cumulative sums of the differences of the forecast errors over time. We plot

$$SSE^{(1,2)}_{h,N} = \sum_{t=t_0}^{N-h} \left[(Y_{t+h} - \hat{Y}^{(1)}_{t+h|t})^2 - (Y_{t+h} - \hat{Y}^{(2)}_{t+h|t})^2 \right]$$

versus N, for $t_0 + h \leq N \leq$ T, where $N - h$ is the endpoint of each truncated time series. Figure 4 shows the 1-step-ahead forecast errors (solid line) and 12-step-ahead forecast errors (dashed line).

Figure 4. Forecast Error Performance

Method C − Method A



Grid Lines at January

A persistently decreasing graph with increasing N (as seen in Figure 4) means that $SSE^{(1,2)}_{h,N}$, the quantity we are adding to the sum, is negative. Negative quantities mean that errors from Model 2 are generally larger than errors from Model 1. We prefer Model 1 because it produces better forecasts (Method C in Figure 4).

A persistently increasing graph with increasing N means that we prefer the forecasts from Model 2.

This diagnostic is not always conclusive. It depends on the forecasts but also on user judgment of how persistent a graph's increase or decrease is.

5. Results

5.1 General Model Adequacy Results

The new visual methods (A, C, and D) were not better than the automatic method (B) according to the model adequacy diagnostics.

We limited our comparisons to include only the series that had different outlier sets for the different methods. Of course, there were similarities among the methods. For 23% of the series (48 of 206) no method identified any outliers for any of the data sets, so we did not include those in any comparisons. For the remaining 158 series, we compared methods, *including Method Z*, to see how many series failed the spectrum diagnostics (residual seasonal effects and trading day effects) and the Ljung-Box Q diagnostic. For further comparisons, excluding Method Z, we used 122 series, eliminating 36 series that had the same outlier set for every identification method (that is, given a particular data set, all identification methods selected the same outliers). Recall that our initial run included data from January 1989 to August 2000 and our final run included data through January 2001.

5.2 Spectrum Diagnostic

The spectrum diagnostics of the regARIMA model residuals did not differ much among methods. Where there were significant differences, no method bettered Method B.

Table 1 shows the number of failures for each method. A Chi-square goodness-of-fit test of the results of the initial run (158 series) showed that at the 95% confidence level, Method B had significantly fewer failures than Methods C and Z. However, in the final run, no methods were significantly different at the 95% level.

The pattern was similar for the 122 series with outlier differences (excluding Method Z). For the initial run, the number of failures under Method C was significantly higher than under Method B (95% level), and Method A failed more often than Method B at the 90% level. For the final run, Method C had significantly more failures than Method B at the 95% level, but this was the only significant difference.

The identification methods were not significantly different in number of trading day spectrum diagnostic failures when including Method Z (158 series). When comparing the 122 series with outlier differences, for the initial run, no method was significantly better than Method B. For the final run there were no significant differences among the methods.

Table 1. Seasonal Spectrum Diagnostic Failures, Number of Series by Method

	A	B	C	D	Z
Initial Run	17	12	20	15	21
158 Series	11%	8%	13%	9%	13%
Final Run	25	21	28	24	25
158 Series	16%	13%	18%	15%	16%
Initial Run	14	9	17	12	-
122 Series	11%	7%	14%	10%	
Final Run	20	16	23	19	-
122 Series	16%	13%	19%	16%	

5.3 Ljung-Box Q Diagnostic

The Ljung-Box Q diagnostic showed more differences among methods than the spectrum, but not in favor of the new visual methods.

Table 2 shows Ljung-Box Q failures. Chi-square goodness-of-fit tests showed that for the initial run, at the 95% confidence level, only Method B had fewer failures than Method Z. Methods A and D had fewer failures than Method Z at the 90% level.

For the final run (158 series), all identification methods were significantly better than Method Z. We expected outlier identification to improve model fit, so the comparison to Method Z in the final run was not surprising. We were disappointed that the initial run did not have this result.

When comparing the 122 series with outlier differences, Chi-square tests showed that for the initial run, at the 95% level, Method C had more failures than Method B. In the final run, Method A had more failures than Method B at the 90% level.

Table 2. Ljung-Box Q Diagnostic Failures, Number of Series by Method

	A	B	C	D	Z
Initial Run	8	6	10	8	14
158 Series	5%	4%	6%	5%	9%
Final Run	8	5	6	6	16
158 Series	5%	3%	4%	4%	10%
Initial Run	6	4	8	6	-
122 Series	5%	4%	7%	5%	
Final Run	6	3	4	4	-
122 Series	5%	2%	3%	3%	

5.4 Outlier T-value Results

In production work, when we add outlier regressors to a model, we are never certain that the regressors will continue to be significant as we add data. We included this concern as one of our comparisons. Our significance criteria were the appropriate critical values as discussed in Section 4.2, Outlier T-values and Additional Outliers.

Even the initial runs of Methods A and D contained some outlier regressors with nonsignificant t-values. There are a few reasons for this result: a) we didn't use a backward elimination step with Methods A and D, b) the difference in variances between what X-12-ARIMA uses initially for the t-values that we saw in our graphs and what it uses for the final model estimation can cause enough difference that the t-values are no longer significant, c) adding even one outlier can change the model fit enough that other observations that initially seemed to be outliers are no longer unusual, and d) our approach to neighboring outliers caused too many AOs in some models.

In the initial run, of the 122 series with different outlier sets, under Method A, 31 had at least one absolute t-value less than the critical value; 13 series failed this test under Method D.

Table 3 shows that in the final run, Method A had the most series (32) that failed this significance test. Of those 32 series, 28 had failed in the initial run (three series failed initially but passed in the final run). Method D had 17 failing series in the final run; 12 of those failed in the initial run. Surprisingly, Method B had only one series with nonsignificant outlier regressors. That series had one outlier that failed by less than 0.005. We expected a few outliers to fall below the critical value over time, so the result for Method B seemed exceptional.

Table 3. Outlier Regressor T-values Failures in Final Run, Number of Series by Method (of 122 Series)

A	B	C	D
32	1	4	17
26%	1%	3%	14%

5.5 Additional Outliers

We also wanted to know how many additional outliers we would identify *in the original span of data* as we added more data. These additional outliers were the result of X-12-ARIMA's automatic identification procedure. We counted newly identified outliers at any month from January 1989 (the start of the data span) up to May 2000 (three months before the endpoint of the initial data span). For some series X-12-ARIMA selected outliers in one run but not in the next. Looking just at the final run, Table 4 shows that by this measure, Method D is the most sufficient method, different from the other methods at the 95% level. It makes sense that Method D would be somewhat lower in additional outliers because it is the combination of two outlier identification methods, but this was a positive result for Method D.

Table 4. Additional Outliers Identified in Final Run, Number of Series by Method (of 122 Series)

A	B	C	D
12	15	18	5
10%	12%	15%	4%

5.6 Forecast Performance Results

Aside from the sufficiency measure shown in Table 4, our diagnostic comparisons indicated that Method B was the best of our methods. Our final comparison was of forecast performance of Methods B and D. We eliminated series that had the same outlier sets, leaving us with 72 series for the initial run and 69 series for the final run. We counted only the series whose models passed the model adequacy diagnostics (spectrum and Ljung-Box Q) under both methods. We counted final runs if they passed the diagnostics even if they had failed during the initial run.

As Table 5 shows, for the initial run we selected Method D as having a better forecast performance more often than Method B. We also found many graphs to be inconclusive. A Chi-square goodness-of-fit test indicates that the preference for Method D is significant at the 95% level. Many of the choices were close. If we consider the 15 questionable choices to be inconclusive, we preferred Method B 10 times and Method D 11 times.

The preferences were not significantly different in the final run, whether or not we included questionable decisions. The initial positive result for Method D didn't hold up over time well enough for us to recommend it.

Table 5. Out-of-Sample Forecast Error Comparison, Preferred Method

	B	D	inconclusive
Initial Run	14	22	19
55 Series	25%	40%	35%
Final Run	13	17	19
49 Series	27%	35%	39%

6. Conclusion

We expected that we would be able to improve on Method B, the automatic outlier identification method, by paying closer attention to the outlier selection and using new identification methods. We expected several examples of improved model performance in the FTD series. We expected especially good results from Method D because it combined the automatic method with our new visual identification method.

Because our visual methods included subjective choices, others who use similar methods may achieve different results. However, the automatic procedure proved itself against our implementation of the visual identification methods. The fact that the procedure requires little additional work from the user makes the decision even easier. Users who have detailed information about a series to know what interventions are needed should use that information. In this research we did not have that kind of knowledge, and we are comfortable concluding that the current automatic identification method is the best of the methods we compared.

Acknowledgments

We thank the staff in FTD—David Dickerson, Amerine Dizon, Neeta Lall, and Melvin McCullough—for their help in getting us data and in giving us their ARIMA models for the series. We are also very grateful to Brian Monsell for giving us software support with X-12-ARIMA.

References

Cleveland, W.S. and S.J. Devlin (1980), "Calendar Effects in Monthly Time Series: Detection by Spectrum Analysis and Graphical Methods," *Journal of the American Statistical Association*, 75: 487-496.

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.

Hood, C.C. (2001), "User's Guide for the X-12-Graph Interactive for PC/Windows, Version 1.2," U.S. Census Bureau, U.S. Department of Commerce.

Ljung, G.M. (1993), "On Outlier Detection in Time Series," *Journal of the Royal Statistical Society*, B, 55: 559-567.

Ljung, G.M. and G.E.P. Box (1978), "On a Measure of Lack of Fit in Time Series Models," *Biometrika*, 65: 297-304.