Comparing the Automatic ARIMA Model Selection Procedures of X-12-ARIMA Versions 0.2 and 0.3

Ayonda M. Dent, Catherine C. Harvill Hood, Kathleen M. McDonald-Johnson and
Roxanne M. Feldpausch
US Census Bureau, ESMPD Room 3112/4, Washington, D.C. 20233-6200

Keywords: seasonal adjustment, time series

1. Background/Motivation

The U.S. Census Bureau is in the process of incorporating an additional ARIMA (Autoregressive Integrated Moving Average) model selection procedure into the X-12-ARIMA software. The automatic modeling procedure in X-12-ARIMA Version 0.2 examines a list of five possible ARIMA models (found in the file x12a.mdl). The new procedure in Version 0.3 has a broader range of models to choose from, and we expect it to fit models to a wider range of series than Version 0.2 (Monsell, 2002). In addition, Version 0.3 has more options than in previous versions.

In this paper, we compare the two automatic modeling procedures. To compare the procedures, we modeled a large group of Census Bureau series including U.S. Imports, Exports, and Retail Sales series. We compared models using diagnostics such as the spectral peaks, Ljung-Box Q (LBQ) statistics, and forecast errors. The goal of our research is to not only determine which procedure performs better, but also give analysts some guidance on the various options available in the new procedure.

Our motivation behind this study was two-fold: 1) to provide users with a documented study of the benefits of the new method, and 2) to give the users some ideas of what to expect from the new automatic modeling procedure once the new version is released.

1.1 Description of the Software

In X-12-ARIMA Version 0.2 and before, the procedure is similar to the automatic modeling procedure in Statistics Canada's X-11-ARIMA (Dagum, 1988). For this study, we used Version 0.2.10 Build 66, which for ease of reference we are calling Version 0.2. The models, by default, in the automatic modeling procedure in X-12-ARIMA Version 0.2 are (0 1 1)(0 1 1), (0 1 2) (0 1 1), (2 1 0)(0 1 1), (0 2 2)(0 1 1), and (2 1 2) (0 1 1). The program chooses the first model for which the diagnostics pass. This allows for models that are more parsimonious because, by default, the simpler models are higher on the list. Alternatively, the user could ask the program to choose the model with the best diagnostics.

By default, the procedure considers the following to be a passing set of model diagnostics. (See X-12-ARIMA Reference Manual):

1. The average absolute percent error (AAPE) of the extrapolated values within the last three years of data is less than 15%,
2. The p-value for the LBQ statistic at lag 24 is greater than 5%, and
3. The sum of the nonseasonal MA parameter estimates (for models with at least one nonseasonal difference) is less than 0.9. (U.S. Census Bureau 2004, p. 51)

If any of the three tests above fail, the model is rejected. Any of the criteria can be altered by specifying an appropriate input argument in the program input (see Section 1.3) For example, setting the value of “fcstlim=25” will allow a series to pass if the AAPE is less than 25%.

X-12-ARIMA Version 0.3 incorporates a new procedure based on the automatic model selection procedure of TRAMO (Time series Regression with ARIMA noise, Missing values and Outliers), a seasonal adjustment program developed by Victor Gómez and Agustín Maravall. TRAMO takes a different approach to model selection, using the Bayesian Information Criterion (BIC) and other model diagnostics to determine the order of differencing and the various MA and AR parameters (Gómez and Maravall, 1997). Models from X-12-ARIMA Version 0.3 will not match exactly the models of TRAMO for every series for several reasons (Monsell, 2002).

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

1.2 Previous Studies

In earlier, undocumented studies, we found we could fit models to a wider range of series with TRAMO than with X-12-ARIMA Version 0.2. For example, in Version 0.2, all the default models are seasonal models, and sometimes we need to fit models to nonseasonal series for outlier detection and trading day adjustment. In one early study, we added all the ARIMA models with p, q≤3; d≤2; P, Q≤1; and D=0,1 where

- p is the nonseasonal order of the autoregressive (AR) component
- q is the nonseasonal order of the moving average (MA) component
- d is the nonseasonal degree of differencing
- P is the seasonal order of the of the AR component
- Q is the seasonal order of the MA component

to the model selection list in X-12-ARIMA Version 0.2. For many series, especially for the nonseasonal series, TRAMO was able to select a much more appropriate model than the extended list in Version 0.2. This led to Version 0.3. For the past few years, we have been investigating the differences between TRAMO and Version 0.3 (See Farooque, Findley, and Hood, 2001).

1.3 Automdl Options

In Version 0.2, the options mostly concern the limits for the tests to see if any given model in the list is acceptable. The most commonly used options are given below:

- **Fcstlim** - sets the threshold for the within-sample forecast error test. The average absolute percent error for the last three years of data must be less than this value for the model to be accepted. The default is fcstlim = 15.
- **Overdiff** - sets the threshold for the sum of the MA parameter estimates in the overdifferencing test. The sum must be less than the limit for the model to be accepted. The default is overdiff = 0.9.
- **Qlim** - sets the acceptance threshold for the p-value of the Ljung-Box Q statistic (at lag 24 for monthly series) for model adequacy. The p-value must be greater than this value for the model to be accepted. The default is qlim = 5, meaning 5 percent or a p-value of 0.05.
- **Method** - specifies whether the automatic modeling procedure will select the first model which passes the three model selection criteria listed above or the model with the best critical values. The default is method = first.

In Version 0.3, the options available for the new automatic modeling procedure differ considerably from the available options in Version 0.2. Though we do not have the same degree of experience with Version 0.3's automatic model identification system as with Version 0.2's, we have studied the following options:

- **Checkmu** - controls whether the automatic model selection procedure will check for the significance of a constant term. The default is checkmu = yes.
- **Diff** - controls the orders of differencing for the ARIMA models, fixing them to the values specified. There is no default; the program will test for various differences based on the maxdiff option listed below.
- **Maxdiff** - specifies the maximum orders of the differencing for the automatic identification of differencing orders. The default is maxdiff = (2 1), meaning the procedure will test for 0, 1, or 2 regular differences and 0 or 1 seasonal difference.
- **Maxorder** - specifies the maximum orders of the ARMA parameters for the automatic identification procedure. The default is maxorder = (2 1), meaning the procedure will test for regular AR and MA parameters up to and including order 2 and seasonal AR and MA parameters up to and including order 1.
- **Mixed** - controls whether models both the AR and MA parameters are considered in the automatic identification procedure. A mixed model is a model with both AR and MA terms in the same model component. The default is mixed = yes.
- **Acceptdefault** - controls whether the default model is accepted based on the Ljung-Box Q-statistic (at lag 24 for monthly series). If the default model is acceptable, no further testing is done. The default is acceptdefault = no.

It is possible to set the options in Version 0.3 to get similar results to Version 0.2. Because Version 0.2 does not check for constant terms, one way to get the results from Version 0.3 to be closer to the results from Version 0.2 would be to disable this option (with checkmu = no) and then to not specify a constant in the regression specification (spec). It would also be possible to set diff = (1 1), meaning that the program would only consider models with one regular and one seasonal difference. Because most of the default models from Version 0.2 have one regular and one seasonal difference, this would help the results to match more closely also. However, we feel that both of these options offer us advantages over the older version. There are several series where a test for a constant term is useful, and where the differencing orders may be different from 1. In our experience, there are several series that do not need one of the differencing terms, and this kind of model is not available in the default model list in Version 0.2.

We have found that many series need AR or MA orders larger than 2, so we generally set maxorder = (3 1) or (4 1), even though, again, this would correspond to models not available by default in Version 0.2.

While testing on the acceptdefault option has been more limited, it may be possible that with acceptdefault=yes, the results could be closer to the results from Version 0.2.

2. Diagnostics

We examined the following diagnostics:

1. Outliers - A large number of outliers can seriously bias the ARIMA (*p, d, q*) x$\left(\mathrm{P}, \mathrm{D}, \mathrm{Q}\right)_s$ process and may indicate a problem with the model. Let W_t represent an outlier-free time series model $\phi_p\left(\beta\right)\Phi_p\left(\beta^S\right)W_t = \theta_q\Theta_Q\left(\beta^W\right)Z_t$ where $W = \nabla^d\nabla_S^D X_t$ and Z_t is a sequence of independent and identically distributed $\mathrm{N}(0, \sigma^2)$ random variables (Box, Jenkins, and Reinsel, 1994).

X-12-ARIMA Version 0.2 and Version 0.3 have the same outlier identification procedure. Therefore, any difference in the outliers identified arises from the ARIMA model selected by each version. In other words, large differences in the ARIMA model residuals could lead to differences in outlier identification. The X-12-ARIMA automatic identification procedure identifies additive outliers, level shifts, and temporary change outliers. Our discussion of outliers will not distinguish between the different outlier types. The automatic outlier identification procedure selects outliers by comparing regressor t-values to a critical value. The length of the series determines X-12-ARIMA's default critical value, but users can set a different critical value. We used the default critical value.

2. Spectral diagnostics for seasonal and trading day peaks - For the purpose of this study, we will define an acceptable model as one without residual seasonal or calendar effects in the seasonally adjusted series or the irregular component. One way to examine these effects is spectral plots. Vertical lines mark seasonal frequencies at *k*/12 cycles/month for $1 \le k \le 6$ and trading day frequencies at .348 and .432 cycles/month (Cleveland and Devlin 1980). Because it is difficult to interpret in economic series, we ignored peaks at 5/12 cycles/month (repeats every 2.4 months). Nor did we examine the seasonal frequency at 6/12 cycles/month or the trading day frequency at .432 cycles/month. (Soukup and Findley, 1999)

3. Ljung-Box Q statistics - Subsequently, to examine the goodness-of-fit we examined the Ljung-Box Q statistics (Ljung and Box 1978). A lag with a p-value less than 0.05 indicates a lack of fit for the model. The seasonal lags for monthly series (12, 24, 36, etc.) are the most important lags. Consequently, models failed the goodness-of-fit test if they exhibited the following characteristics:

a. Lag 12 failed
b. More than six lags failed from lag 1 to 12 (first seasonal lag)
c. More than 12 lags failed from lag 1 to 24 (second seasonal lag)

(See also McDonald-Johnson, Hood, Feldpausch, 2004).

4. Forecast Errors - We also used the average absolute percent within-sample forecast error to compare model residuals. A smaller forecast error means the model has better forecasting performance. This refers to the model's ability to reproduce data that are already known.

3. Methods

We began our study with 311 data series from U.S. Retail, Imports, and Exports series. However, since the ARIMA models identified in the automatic modeling selection procedure in Version 0.2 all have seasonal components by default, we eliminated the nonseasonal series. As a result, we removed four series, leaving 307. Then we created basic input specification files using the Windows® Interface to X-12-ARIMA (Feldpausch, 2003) to create basic input specification files for the series. These specifications files were run twice for every series, once in Version 0.2 (build 66, compiled July 27, 2004) and in Version 0.3 (build 139, compiled

July 14, 2004). We ran each specification file using default settings.

4. Results

Of the 307 series, both versions of X-12-ARIMA chose the same models for 170 series (55.4%). Table 1 shows a breakdown by series. With the exception of one series, when they did agree, the model Version 0.2 selected was (0 1 1)(0 1 1). As stated in section 1.1, this is the first model by default that Version 0.2 is set to choose.

Table 1. Frequency Same Model was Chosen by Series

Series	Frequency
Retail	17/43 (39.5%)
Imports	74/122 (60.6%)
Exports	79/142 (55.6%)

Table 2 shows a frequency summary of the models chosen by Version 0.2. In contrast, Version 0.3 selected a variety of models, including several mixed models. (See Table 3 for a summary of the top five models chosen by Version 0.3.)

Table 2. Version 0.2 Models Chosen Using Default Parameters

Model	Frequency
(0 1 1)(0 1 1)	270 (87.9%)
(0 1 2)(0 1 1)	24 (7.8%)
(2 1 0)(0 1 1)	9 (2.9%)
(0 2 2)(0 1 1)	0
(2 1 2)(0 1 1)	4 (1.3%)

Table 3. Version 0.3 Top Five Models Chosen

Model	Frequency
(0 1 1)(0 1 1)	185 (60.3%)
(0 1 0)(0 1 1)	20 (6.5%)
(0 1 2)(0 1 1)	10 (3.3%)
(1 1 0)(0 1 1) *	9 (2.9%)
(2 1 1)(0 1 1) *	9 (2.9%)

*model not available in Version 0.2

Simpler models where ARIMA factors have either an AR or MA part can be more advantageous. Thus, at the Census Bureau such models are often preferred to mixed models. (See McDonald-Johnson, Hood, Feldpausch, 2004) Version 0.2 selected mixed models 1.3% (4) of the time while Version 0.3 selected them 10.7% (33) of the time.

4.1 Outliers

We did not find a major difference in the number of outliers identified by X-12-ARIMA Version 0.2 and Version 0.3. Table 4 shows the average and total number of outliers for each version. These numbers alone do not tell us which version does a better job of model selection.

Table 4. Outlier Distribution

Series		**0.2**	**0.3**
Retail	Avg	2	2
	Total	84	80
Imports	Avg	1	1
	Total	145	146
Exports	Avg	1	1
	Total	120	121

For series with considerable differences, we observed the following:

1. Version 0.2 would select a simpler model with a greater number of outliers for the series. At the same time, Version 0.3 often selected mixed models, identified fewer outliers, and appeared more acceptable. For example, for one Retail series, Version 0.2 selected the ARIMA model (0 1 2)(0 1 1) and identified ten outliers. Version 0.3 selected (2 1 1)(0 1 1) and identified three outliers.

2. In addition, for the series that Version 0.3 selected fewer outliers, further investigation of the diagnostics files showed the additional outliers identified by Version 0.2 were chosen as *almost outliers* in Version 0.3. That is, X-12-ARIMA produces a list of outliers whose t values are within 0.5 of the critical value or that were identified as outliers but removed in the backward elimination step of the outlier identification procedure. We call these *almost outliers*. For one series, Exports of Crude Oil, Version 0.3 chose 14 outliers and Version 0.2 chose 19 outliers. In the diagnostics file, three of the outliers listed as almost outliers were identified as outliers in Version 0.2.

4.2 Spectral Diagnostics

Overall, the spectral diagnostics did not differ much between the two versions. The one noticeable difference was the number of trading day peaks in the spectrum of the seasonally adjusted series for Version 0.3 nearly doubled that of Version 0.2.

Of the 43 Retail series, both versions of X-12-ARIMA produced similar spectral diagnostics—yielded the same types of peaks or no peaks at all—for 37 (86%) series. Table 5 shows the number of series with visually significant seasonal and trading day peaks in the spectrum of the seasonally adjusted series. Table 6 provides details of the series that produced different spectral results. Note that S1-S4 corresponds to seasonal frequencies at 1/12, 2/12, 3/12, 4/12 cycles/month, respectively and T1 corresponds to trading day frequency .348 cycles/month.

Table 5. Retail Series, Number of Visually Significant Peaks

Peaks	Version 0.2	Version 0.3
No peaks	23 (53.5%)	23 (53.5%)
S1-S4	19 (44.2%)	15 (34.9%)
T1	2 (4.7%)	6 (14.0%)

Table 6. Retail Series, Spectral Differences

Series	Version 0.2	Version 0.3
series1	S4	none
series2	none	T1
series3	none	T1
series4	S1	T1
series5	S3	none

Similar patterns between Versions 0.2 and 0.3 emerge for 90.1% of the Import series. Of the 142 series, 128 had comparable spectral diagnostics for both versions. Likewise, of the 122 Export series, 111 (90.9%) had similar results. See Tables 7 and 8 for more details.

Table 7. Imports Series, Number of Visually Significant Peaks

Peaks	Version 0.2	Version 0.3
No peaks	109 (76.8%)	102 (71.8%)
S1-S4	26 (18.3%)	27 (19%)
T1	7 (4.9%)	13 (9.2%)

Table 8. Exports Series – Number of Visually Significant Peaks

Peaks	Version 0.2	Version 0.3
No peaks	96 (78.7%)	82 (67.2%)
S1-S4	20 (16.4%)	25 (20.5%)
T1	12 (8.5%)	19 (15.6%)

4.3 Ljung-Box Q

No major differences existed in the goodness-of-fit diagnostics. Of the 307 series, 30 series from Version 0.2 and 18 series from Version 0.3 failed this diagnostic. In the cases where the series in Version 0.3 passed and same series in Version 0.2 failed, Version 0.3 frequently selected a mixed model while Version 0.2 selected a simpler model.

4.4 Forecast Errors

The final diagnostic we compared was the forecast performance. There were no major differences in the average absolute percent forecast errors. Figures 1 through 3 are plots of the absolute difference of the within-sample forecast errors for each series. Minor dissimilarities exist, but overall both versions produce similar patterns.

Figure 1. Within-Sample Forecast Errors Retail

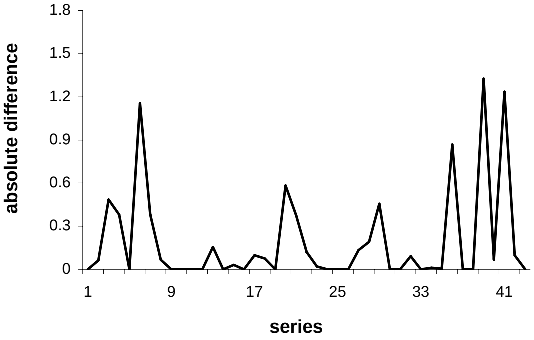


Figure 2. Within-Sample Forecast Errors Imports

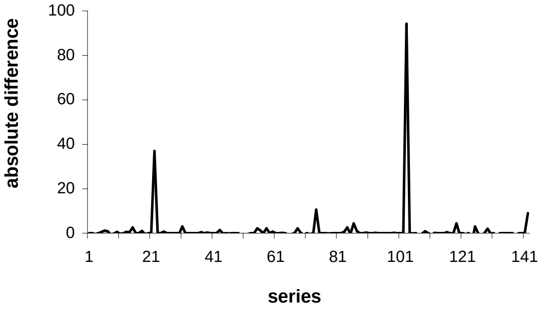
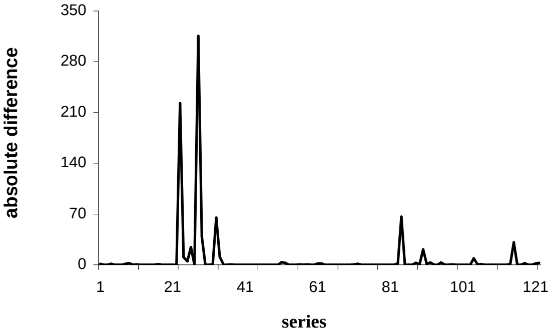


Figure 3. Within-Sample Forecast Errors Exports



5. Conclusion

We could not conclude that one version produced better diagnostics than the other version. Where differences do exist, they are not major. However, one underlying assumption that could affect the user's decision when choosing between the two versions is whether the data are seasonal or nonseasonal. Because Version 0.2 assumes all data are seasonal, Version 0.3 has the advantage when the data the user is working with may not be seasonal. If prior knowledge of this sort were not known, Version 0.3 would require a limited amount of additional work by the user. Moreover, X-12-ARIMA Version 0.3 incorporates the automatic modeling procedure found in Version 0.2.

6. Future Study

For future investigation, we would like to use simulated series where we would know what the underlying model should be.

Acknowledgements

We would like to thank Brian Monsell, Thomas Trimbur, and Rita Petroni for their valuable comments and suggestions.

References

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994), *Time Series Analysis: Forecasting & Control*, 3rd Edition, Prentice Hall.

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

Dagum, E. B. (1988), "The X-11-ARIMA/88 Seasonal Adjustment Method – Foundations and Users' Manual," Statistics Canada.

Farooque, G.M., Hood, C.C., and Findley, D.F. (2001) "Comparing the Automatic ARIMA–Model Selection Procedures of TRAMO and X-12-ARIMA Version 0.3," 2001 *Proceedings of the American Statistical Association*.

Feldpausch, R. (2003), "Windows® Interface to X-12-ARIMA," U.S. Census Bureau, U.S. Department of Commerce.

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

Ljung, G.M. and G.E.P. Box (1978), "On a Measure of Lack of Fit in Time Series Models," *Biometrika*, 65, 2, pp. 297-3-3.

McDonald-Johnson, K.M., Hood, C.H., Feldpausch, R. (2004), "Model Simplification After the Automatic Modeling Procedure of X-12-ARIMA Version 0.3," *Proceedings of the American Statistical Association*.

Monsell, B.C. (2002) "An Update on the Development of the X-12-ARIMA Seasonal Adjustment Program," *Proceedings of the 3rd International Symposium on Frontiers of Time Series Modeling*, pp. 1–11.

Soukup, R. J. and D. F. Findley (1999). "On the Spectrum Diagnostics Used by X-12-ARIMA to Indicate the Presence of Trading Day Effects After Modeling or Adjustment," *Proceedings of the American Statistical Association*

U.S. Census Bureau (2004), X-12-ARIMA Reference Manual, Version 0.3, U.S. Census Bureau, U.S. Department of Commerce.