# Improving the Automatic RegARIMA Model Selection Procedures of X-12-ARIMA Version 0.3

**Kathleen M. McDonald-Johnson, Thuy Trang T. Nguyen, Catherine C. Hood, and Brian C. Monsell**
**U.S. Census Bureau, ESMPD Room 3110/4, Washington, D.C. 20233-6200**

**Key Words:** **seasonal adjustment; regression model with ARIMA errors; out-of-sample forecast error comparisons; F-adjusted Akaike's Information Criterion**

## 1. Summary

The U.S. Census Bureau has enhanced the X-12-ARIMA seasonal adjustment program by incorporating an improved automatic regARIMA model (regression model with ARIMA errors) selection procedure. Currently this procedure is available only in test version 0.3 of X-12-ARIMA, but it will be released in a future version of the program. It is based on the automatic model selection procedure of TRAMO, an ARIMA-modeling software package developed by Víctor Gómez and Agustín Maravall (Gómez and Maravall 1997). The procedure of X-12-ARIMA differs from that of TRAMO in several ways, related mainly to parameter and likelihood calculation and to outlier identification. We looked at ways to determine presence of trading day (TD), Easter, and outlier effects to possibly improve the ARIMA model chosen by X-12-ARIMA. We compared models using diagnostics such as out-of-sample forecast error graphs, spectral analysis, Ljung-Box Q statistics, and under certain circumstances, the Hannan-Quinn statistic.

We concluded that we need further research to determine the best procedure for selecting TD and Easter regressors. We have changed the automatic modeling procedure. The F-adjusted Akaike's Information Criterion (AICC) is now the primary selection tool, but the program also uses the regression t-values to eliminate nonsignificant regressors. We could not determine whether changing the outlier critical value during the automatic model selection can improve the final model.

## 2. Background

The U.S. Census Bureau continues to improve X-12-ARIMA, the most recent seasonal adjustment program in the X-11 line (Findley, Monsell, Bell, Otto, and Chen 1998). X-12-ARIMA follows X-11, developed at the U.S. Census Bureau (Shiskin, Young, and Musgrave 1967), and X-11-ARIMA and its further developments from Statistics Canada (Dagum 1988).

One major improvement of X-12-ARIMA over X-11 is the use of regARIMA models to estimate calendar effects or outlier effects with predefined or user-defined regressors. X-12-ARIMA uses regARIMA models to remove effects such as TD, moving holidays, and outliers before performing seasonal adjustment. In addition, forecast extensions from the models can improve the X-11 filter result at the end of the series. Improving regARIMA model selection should improve the quality of the prior adjustments and the forecast performance, leading to a better quality seasonal adjustment result.

X-12-ARIMA can determine various regARIMA options with several automatic procedures:

- choice of series transformation (log function or no transformation),
- selection of regression effects such as TD, Easter, and outliers, and
- determination of ARIMA model (including the trend constant regressor if the absolute value of the regression t-value is greater than 1.96).

Details of the procedure can be found in Monsell (2002).

In this paper we discuss methods of selecting TD and Easter regression effects. We also consider how outlier identification affects the automatic modeling procedure.

The usual flow TD regression (the primary TD effect used for this paper) estimates the effect on the series value from the weekday composition of the measurement period. For example, in any given month, each day of the week occurs at least four times. Days that occur five times may affect the value for that month. If activity is strong on Saturdays, a month with five Saturdays may have a larger value than a month with only four Saturdays. X-12-ARIMA estimates six regression variables, with a seventh variable constrained by the sum of the others. A stock TD variable is also available (Findley et al. 1998).

The Easter regression estimates holiday activity that starts $\omega$ days before Easter and ends the day before Easter. It is denoted by Easter[$\omega$] where $\omega$ can range from 1 to 25. When testing for Easter effects, given an ARIMA model

---

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

and no specific ω, X-12-ARIMA estimates three Easter regressors: Easter[1], signifying an effect occurring on the day before Easter; Easter[8], signifying a week-long effect that starts eight days before Easter; and Easter[15], signifying a two-week effect (U.S. Census Bureau 2002, p. 108). We consider these three regressors in this paper.

## 3. Regression Selection Tools

### 3.1 Motivation for Regressor Selection Study

For many of the test comparisons described above, X-12-ARIMA uses the AICC (Findley et al. 1998). This criterion differs from the usual AIC statistic because it includes a correction for the length of the series. The U.S. Census Bureau's Time Series Staff generally regards AICC as the best tool for these comparisons, and in this study we compared one approach using AICC to another approach using significance tests.

When users specify an ARIMA model and request an AICC test for the presence of a TD or Easter effect (the input specification option is *aictest*), X-12-ARIMA calculates AICC values for the model with and without the effect and chooses the regression model with the minimum AICC. However, until recently, under automatic modeling, when users requested a test for presence of a TD or Easter effect, X-12-ARIMA performed a t test to determine the presence of the effect. For instance, when testing for a TD effect, the program fit the default ARIMA model (usually the airline model) with the TD regression. If at least one estimated t-value for the TD regressors was greater than or equal to 1.96 in absolute value, then the program included the TD effect. This procedure was the same as TRAMO's regressor selection method, although it performed differently under testing (Farooque, Findley, and Hood 2001; Hood 2002).

Because one significant TD t-value can occur even when the combined effect is not significant (as measured by a chi-square statistic), the Time Series Staff decided to change the regressor selection method.

It would be consistent for X-12-ARIMA to use the minimum AICC as a regressor selection tool under automatic modeling as it does when the user specifies the ARIMA model. We were concerned that computing AICC would be considerably slower than computing t or chi-square statistics. We would prefer to use a faster, simpler significance test if it performed as well as AICC.

AICC and hypothesis tests have different objectives, and AICC results correspond to a higher alpha significance level (perhaps 0.15 or 0.20) rather than the usual 0.05 or 0.01 test levels (W. R. Bell, personal communication, September 11, 2002). We recognize this as a potential weakness of our study, but here we share what we learned in this attempt to improve the modeling procedure.

### 3.2 Methods and Results of Regressor Selection

For the regressor selection study, we devised two methods, A and B. For Method A we used minimum AICC to select regressors. For Method B we used significance tests. We assessed the methods in three stages: (1) TD assessment, (2) Easter assessment, and (3) automatic modeling assessment.

The automatic modeling assessment included additional steps, but all three assessments began with certain input settings described below. For all series we performed the same regressor selection method, but we assessed the selections separately. That is, we did not assess the Easter regression selection for the TD assessment series, and we did not assess the TD regression selection for the Easter assessment series. We did not perform seasonal adjustment during any of the runs.

For both Method A and Method B, we set the model options to simulate the usual automatic modeling settings:

- (0 1 1)(0 1 1) ARIMA model (airline model, the usual default model for automatic modeling),
- automatic outlier identification procedure to identify additive outliers (AOs), level shifts (LSs), and temporary changes (TCs), and
- automatic transformation choice to determine whether or not to take the log of the series (based on AICC, with a slight bias toward log transformation).

*Method A:* Within a single exterior run, X-12-ARIMA fit the model with and without the TD and Easter regressors, computing the AICC for each model. (The fit is sequential. If the AICC favored the TD effect, then the Easter regressor test was fit with the TD regression. If the AICC favored no TD effect, then the Easter regressor test was fit alone.) Note that X-12-ARIMA identified outliers only after selecting the TD and Easter regressors, so outlier choices did not affect the regressor selection.

*Method B:* We fit the model with four regressors simultaneously: TD, Easter[1], Easter[8], and Easter[15]. If X-12-ARIMA identified outliers, then it added them to the final regARIMA model. We used different test statistics for the different regressors. For the TD regression, if the p-value of the chi-square statistic was less than 0.01, we accepted the TD effect. For the Easter regression, if the absolute value of any Easter coefficient t-value was greater than 1.96, then we accepted the Easter regressor with the greatest absolute t-value.

The only difference between the input files for Methods A and B was the regression specification.

*Method A:*

```
regression {aictest = (td Easter)}
```

*Method B:*

```
regression {variables = (td Easter[1]
   Easter[8] Easter[15]}
```

Because the AICC is closely related to the likelihood ratio test, for large samples, it is equivalent to the chi-square test, although in our comparisons, the significance levels were not the same (W. R. Bell, personal communication, September 11, 2002). Perhaps the characteristics of Method B that best differentiate it from Method A are 1) the number of TD and Easter regressors fit at one time and 2) the impact of outlier identification on significance of the TD and Easter regressors.

For the TD and Easter assessments, we fit the models to differing spans of time. We completed four runs with different spans, each time removing one year (12 months) from the beginning of the series.

### 3.2.a TD Assessment

The TD assessment involved 141 U.S. import series:

- spans started at January 1989, 1990, 1991, and 1992
- spans ended at August 2000 each time

We compared the TD regression selection results of Methods A and B to a previous study (Hood 2000). This study differed from our research because it involved ARIMA models that were reviewed and selected for each series, and our research involved only the airline model. Also, this study involved diagnostics such as AICC, regression chi-square statistics, spectral plots, and when necessary, out-of-sample forecast error graphs (Findley et al. 1998, sec. 4.3.2). We were confident that the previous study made accurate decisions with regard to TD selection for the series.

For each series we preferred the method with the greater number of spans that agreed with the previous study. We then compared how many times we preferred each method. Table 1 shows these comparison results.

Of the 141 series, Methods A and B always made the same decisions for 110 series. Ninety of those series agreed with the previous study for every span, 15 decisions disagreed with the previous study for every span, and five differed by span. Another series was a tie between methods – each method agreed with the previous study for one span although not the same span. If we had used a different significance level for the chi-square statistic, we may have seen even greater agreement.

Of the remaining 30 series, we preferred Method A 19 times and Method B 11 times. We tested the significance of this result under the null hypothesis that the probability of preferring Method A (or Method B) is 0.5. Under the binomial distribution, the probability that we would prefer Method A 19 or more times in 30 comparisons is 0.1002, so the result is not significant at the 95% level, but it is significant at approximately the 90% level. We could not conclude that the two methods have different probabilities of preference, and yet we were not convinced that Method B was performing as well as Method A.

**Table 1. Trading Day Assessment Results**

| Preferred Method | Number of Series | Percent of Total |
|---|---|---|
| Neither | 111 | 78.7% |
| A | 19 | 13.5% |
| B | 11 | 7.8% |
| Total | 141 | 100.0% |

### 3.2.b Easter Assessment

The Easter assessment involved 46 retail sales series:

- spans started at January 1987, 1988, 1989, and 1990
- spans ended at December 1998 each time

We evaluated the results by comparing the two methods directly. We considered agreement to mean that the two methods made the same decisions for at least three of the four spans. We did not constrain agreement by duration of the Easter effect (ω value). Table 2 shows the comparison results.

**Table 2. Easter Assessment Results**

| Methods A and B | Number of Series | Percent of Total |
|---|---|---|
| Agree: | 33 | 71.7% |
| Easter Effect | 18 | 39.1% |
| No Easter Effect | 15 | 32.6% |
| Disagree: | 13 | 28.3% |
| Method A Easter | 10 | 21.7% |
| Method B Easter | 3 | 6.5% |
| Total | 46 | 100.0% |

Of the 46 series, Methods A and B agreed 33 times and disagreed 13 times. Method A identified an Easter effect for 28 series, including 10 for which Method B did not identify an Easter effect. Method B identified an Easter effect for 21 series, including three for which Method A did not identify an effect.

We tested the null hypothesis that the probability that the two methods will agree is 0.9. Under the binomial distribution, the probability of 33 or fewer agreements is 0.0004, so we reject the null hypothesis. Method B did not match Method A as closely as we would have liked, but we realize that using a different t-value likely would have given more similar results.

### 3.2.c Automatic Modeling Assessment

The automatic modeling assessment was an extension of Methods A and B as described above. We studied 34 monthly series: 27 U.S. import and export series and seven retail sales series. The import and export series had been previously identified as difficult to model using the automatic modeling procedure of X-12-ARIMA (Hood 2002).

We performed the Method A and B regressor selections and hardcoded these results:

- TD effect,
- Easter effect,
- outlier effect(s), and
- transformation choice.

We then ran the automatic modeling procedure of X-12-ARIMA. We hardcoded the final model in addition to the model effects listed above. We then ran X-12-ARIMA with that specified model and collected diagnostics.

For example, for one series, the procedure produced the following regARIMA models. (Notation for outliers is type (AO, LS, or TC) followed by the date (year.mon), that is, TC1996.Mar indicates a TC in March 1996.)

*Method A:*

```
ARIMA {model = (0 1 1)(1 0 0)}
regression {variables = (Easter[1]
   TC1996.Mar LS1992.Jan)}
```

*Method B:*

```
ARIMA {model = (0 1 1)(0 1 1)}
regression {variables = (Easter[15]
   AO1996.Mar LS1992.Jan)}
```

We compared the models from the different methods using the scoring system described in Farooque et al. (2001). The system assigns weighted penalties to the models based on standard model diagnostics:

- Ljung-Box Q (Ljung and Box 1978)
- Spectrum of the regARIMA model residuals (Cleveland and Devlin 1980, Soukup and Findley 1999)
- Hannan-Quinn statistic (Hannan and Quinn 1979)
- Mean of the squared out-of-sample forecast error at leads 1 and 12

We prefer the less-penalized model (Method A in the example shown above).

Table 3 shows the comparison results. Of the 34 series, Methods A and B selected the same model 18 times. Of the remaining 16 series, we preferred Method A's model 11 times and Method B's model five times. We tested the significance of this result under the null hypothesis that the probability of preferring Method A (or Method B) is 0.5. Under the binomial distribution, the probability that we would prefer Method A 11 or more times in 16 comparisons is 0.1051, so the result is not significant at the 95% level. We could not conclude that the two methods have different probabilities of preference, but as with the TD assessment, we were not confident in choosing Method B over Method A.

**Table 3. Automatic Modeling Assessment Results**

| Preferred Method | Number of Series | Percent of Total |
|---|---|---|
| Neither | 18 | 52.9% |
| A | 11 | 32.4% |
| B | 5 | 14.7% |
| Total | 34 | 100.0% |

### 3.3 Conclusions of Regressor Selection Study

The results of the Regressor Selection study were not strongly conclusive. For the TD and automatic modeling assessments, the probability of preference was not significantly different from 0.5 for the two methods, but we were not convinced that the significance tests we used would give us the appropriate regressors.

X-12-ARIMA version 0.3 did change after we completed this research. Now minimum AICC is the initial criterion for selecting TD or Easter regressors during the automatic modeling procedure. Because the selection is first made using the default ARIMA model, there is a second AICC test if the automatic modeling procedure chooses a different model. In addition, to reduce the number of false positive results, X-12-ARIMA removes TD or Easter regression effects that are not significant, that is, if the absolute values of the regressor t-values are less than 1.96 (one t-value for Easter, but all seven t-values in the case of TD). We have implemented the AICC test together with the significance test. This solution has increased the program's computations, rather than reducing them. We would like to continue studying the regressor selection procedure keeping in mind the relationship between AICC and the significance tests that are available.

Future study will involve simulated series so that we can see how well the program identifies known TD and Easter effects. We also will look at the implications for short series.

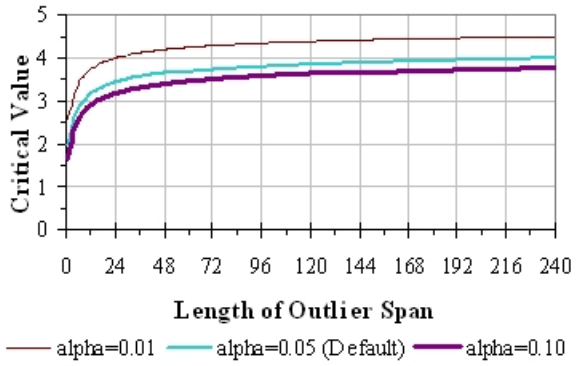
To compare similar significance levels, we may perform similar tests but raise the alpha level for the chi-square or t statistic. As an alternative, we may try requiring a minimum difference in the AICC values before including the regressor(s).

## 4. Differing Outlier Critical Values

Our second study of the automatic modeling procedure involved changing the level of outlier detection. X-12-ARIMA identifies outliers by comparing the regression t-values for different outlier types (AOs, LSs, and TCs) to a preset critical value. The default critical value depends on the length of the span being tested and is set at a 95% confidence level (critical alpha = 0.05). Based on formulas found in Ljung (1993) with interpolation for short spans, the critical value increases monotonically with span length. Users can set a different critical value specifying a value or an alpha level. We wanted to know if systematically lowering or raising the critical value would improve model selection.

Figure 1 demonstrates the relationship between critical value and outlier span length for three alpha values: 0.01 (99% confidence, raising the critical value from the default), 0.05 (95% confidence, the default level), and 0.10 (90% confidence, lowering the critical value from the default).

**Figure 1. Critical Value by Alpha Level**




The study of outlier critical values encompassed two viewpoints. One viewpoint was that allowing the program to detect more outliers during the automatic modeling procedure would improve its ability to determine the underlying process in the series. The opposite viewpoint was that the program tends to select too many outliers, and it would select the best model if it allowed for only the most significant outliers. We looked for changes when we lowered or raised the critical value for outlier identification when running the automatic modeling procedure.

### 4.1 Methods and Results of Differing Outlier Critical Values

Our study included 63 series: 36 construction series (including seven stock series) and the same 27 difficult-to-model U.S. import and export series used in the automatic modeling assessment of the AICC test. We used model spans matching what is used in production runs, and we performed outlier identification on the full model span. The spans ranged from 104 months to 248 months.

Before running the automatic modeling procedure, we set the TD regression adjustment. We did not model Easter effects for any of these series. We based TD decisions on the current production ARIMA model, using the seven-day flow TD effect for flow series and an end-of-the-month stock TD effect for stock series. We used AICC test results and the spectrum of the regARIMA model residuals to select the appropriate TD effects.

After hardcoding the TD regression decisions, we ran the automatic modeling procedure with the automatic transformation choice. We did not run any seasonal adjustment specification. We completed three sets of runs, each with a different critical alpha value: 0.01, 0.05, and 0.10.

If a series' final ARIMA model changed after we raised or lowered the critical value, we hardcoded the transformation choice and new model and refit the model, this time using the default outlier critical value. (We did not refit the models from the original default-level runs.) We then compared models using the out-of-sample forecast error graph and spectrum of the model residuals.

Of the 63 series, X-12-ARIMA lowered the critical value for at least one run for five series. Because they did not represent the systematic change that we were studying, we eliminated them from further comparisons. Only 10 of the remaining 58 series had different models selected when we changed the critical alpha value. One series had a different model for each alpha value, so we compared the three and chose one preferred alpha level. Another series had inconclusive diagnostics, so we had no preference.

We had alpha preferences for only nine series. As shown in Table 4, we preferred the default alpha level more often than the other levels, but these results were not conclusive.

**Table 4. Automatic Modeling Assessment Results**

| Preferred Alpha | Number of Series | Percent of Total |
|---|---|---|
| None | 49 | 84.5% |
| 0.01 | 3 | 5.2% |
| 0.05 | 5 | 8.6% |
| 0.1 | 1 | 1.7% |
| Total | 58 | 100.0% |

### 4.2 Conclusions for Differing Outlier Critical Values

We could not conclude whether or not changing the critical alpha value improves model selection. It may be that outlier selection does not strongly affect model selection. Perhaps a larger sample of series would give a clearer result.

We are not planning to continue this part of the research because we uncovered no evidence that systematically changing the alpha value would improve the modeling procedure.

### Acknowledgments

We would like to thank William R. Bell and Michael Z. Shimberg of the U.S. Census Bureau for their valuable comments and suggestions.

### 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.

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

Farooque, G. M., D. F. Findley and C. C. Hood (2001), "Using the Automatic ARIMA Selection Procedures of TRAMO and X-12-ARIMA 0.3," 2001 Proceedings of the American Statistical Association, Business and Economics Section [CD-ROM], Alexandria, VA: American Statistical Association.

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.

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

Hannan E. J. and B. G. Quinn (1979), "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society*, B, 41: 190-195.

Hood, C. C. (2000), "Results From the Quality Review of X-12-ARIMA Input Files and Recommendations for Changes to Seasonal Adjustment Options," U.S. Census Bureau Internal Memorandum.

——— (2002), "Comparing the Automatic ARIMA Model Selection Procedures of TRAMO and X-12-ARIMA Version 0.3 and the Seasonal Adjustments of SEATS and X-12-ARIMA," unpublished work presented at the Eurostat Working Group on Seasonal Adjustment Meeting, Luxembourg, April 2002.

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.

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 in Time Series Modeling*, Institute of Statistical Mathematics, Tokyo, pp. 1-11.

Shiskin, J., A. H. Young, and J. C. Musgrave (1967), "The X-11 Variant of the Census Method II Seasonal Adjustment Program," Technical Paper No. 15, U.S. Census Bureau, U.S. Department of Commerce.

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," *American Statistical Association 1999 Proceedings of the Business and Economics Section*, pp. 144-149.

U.S. Census Bureau (2002), *X-12-ARIMA Reference Manual, Version 0.2.10*, U.S. Census Bureau, U.S. Department of Commerce.