Model Simplification After the Automatic Modeling Procedure of X-12-ARIMA Version 0.3

Kathleen M. McDonald-Johnson, Catherine C. H. Hood, and Roxanne M. Feldpausch
U.S. Census Bureau, ESMPD Room 3112/4, Washington, D.C. 20233-6200

Key Words: **regARIMA model; seasonal adjustment; time series**

1. Summary

When Statistics Canada released X-11-ARIMA (Dagum 1980, 1988), the improvements to X-11 (Shiskin, Young, and Musgrave 1967) included an automatic modeling procedure to help users take advantage of new program features. X-12-ARIMA (Findley, Monsell, Bell, Otto, and Chen 1998), the most recent program in the X-11 line, retains the X-11-ARIMA automatic modeling procedure, but X-12-ARIMA Version 0.3 includes an additional automatic modeling procedure based on the procedure found in TRAMO (Gómez and Maravall 1997). At the U.S. Census Bureau, we use automatic modeling tools to identify regARIMA models (regression models with ARIMA errors) for forecast extension and estimating regression effects before running the X-11 method of seasonal adjustment. We are concerned that the automatic procedure may identify (1) mixed models that are generally too complicated for concurrent adjustment production at the Census Bureau or (2) models that are susceptible to convergence problems as indicated by coefficient values. We created a program to check automatically-identified models for these concerns and produce X-12-ARIMA input files using simplified versions of the automatically-chosen models. So far, our model-simplification procedure is not completely automated, but we were able to choose simplified models without excessive human intervention.

We compared adjustments using the automatically-identified models, our simplified models, and the airline model, ARIMA (0 1 1)(0 1 1) (Box, Jenkins, and Reinsel 1994). We compared models and the resulting adjustments using goodness-of-fit diagnostics, spectrum of the model residuals, within-sample and out-of-sample forecasts, and revision history diagnostics. According to these diagnostics, results from simplified models were comparable to results from the automatically-identified models.

2. Background and Motivation

Since the U.S. Census Bureau released X-11 (Shiskin et al. 1967), there have been continual updates to the program such as X-11-ARIMA and its further developments from Statistics Canada (Dagum 1980, 1988) and X-12-ARIMA developed at the Census Bureau (Findley et al. 1998).

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.

One major improvement made available in X-11-ARIMA is the use of forecast extension. Treating regARIMA model forecasts like real data so that X-12-ARIMA can apply symmetric X-11 filters at the end of the series can reduce revisions of the seasonally adjusted series. Improving regARIMA model selection should improve the forecast performance, leading to a better-quality seasonal adjustment.

X-11-ARIMA includes an automatic modeling procedure that chooses ARIMA models from a list. In addition to that method, X-12-ARIMA Version 0.3 includes a more flexible automatic modeling procedure based on the method found in TRAMO (Gómez and Maravall 1997).

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

- Choice of series transformation (log function or no transformation),
- Determination of ARIMA model (checking for orders of differencing and significance of ARMA coefficients), and
- Selection of regression effects such as trend constant, trading-day, Easter, and outlier effects.

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

Ongoing research at the Census Bureau (Hood and Findley 1999, Farooque, Findley, and Hood 2001) has compared the results of the automatic modeling procedures of X-11-ARIMA, X-12-ARIMA Version 0.3, and TRAMO using model-fit diagnostics (Ljung-Box Q and AICC when appropriate), number of outliers (too many outliers may indicate problems with a model), and residual effects as shown by peaks at seasonal and trading-day frequencies in the spectrum of the model residuals. See Soukup and Findley (1999) for a description of the spectrum diagnostic.

The research also showed that some automatically-chosen models from previous builds of X-12-ARIMA Version 0.3 were misspecified, mostly due to unit roots, meaning that the sum of the AR coefficients or the sum of the MA coefficients was almost one in absolute value. AR coefficients that sum to almost one indicate possible underdifferencing. MA coefficients that sum to almost one indicate possible overdifferencing. For example, an ARIMA (1 0 1) model with an AR coefficient of 1.0 would be better specified as an ARIMA (0 1 1) model.

The new automatic modeling procedure now includes a final unit root test (U.S. Census Bureau 2004), and AR-model unit roots are not likely to occur. More often, unit-root concerns involve seasonal MA models.

The current test for nonseasonal MA coefficients allows coefficients to be very close to one (up to 0.999). SEATS has a lower limit, fixing MA coefficients at 0.98 for estimates over 0.98.

These limits seem high given the variability of coefficient values from month to month. We chose a cutoff of 0.96 for all coefficient types.

In addition, we are concerned about the automatic choice of mixed models – models with nonseasonal AR and MA terms or seasonal AR and MA terms. It is possible for AR terms and MA terms to cancel each other's effects, so often we can replace mixed models by more simple models.

Cancellation between AR and MA terms occurs with terms that are in the same model component. A model with AR and MA terms in separate components is not a mixed model. For example, the model (2 1 2)(0 1 1) is mixed because the nonseasonal component has AR and MA terms. The model (2 1 0)(0 1 1) is not a mixed model because the nonseasonal AR terms will not cancel the effect of the seasonal MA term.

Some background research reinforced the idea that automatic modeling results may be misleading. After simulating 500 airline-model series, we ran the automatic modeling procedure for four subspans of each original series. The automatic procedure chose the airline model for all four data spans for only 35% of the series. For 5% of the series, the automatic procedure did not choose a seasonal difference for at least one of the spans.

The Census Bureau is committed to having an automatic modeling procedure that is simple enough for inexperienced modelers to use yet flexible enough for more sophisticated users. In addition, the procedure must select appropriate models for users of all levels.

For production seasonal adjustment at the Census Bureau, we specify the regARIMA model, but we reestimate model coefficients as we add data points. The models must converge each month or quarter with no changes to the model settings and no interventions from analysts. Rules of thumb for model selection help keep the models simple and likely to converge.

We often choose simple models to fit well-behaved series. It may be even more important to use simple models for poorly-behaved series, yet those are the series that automatic procedures tend to fit with complex models.

In this context, a simple model is one that is not mixed and has no unit root problems. We describe these as production-ready models because they fit our criteria for production use.

Note that production-ready status as described here is based on seasonal adjustment circumstances at the U.S. Census Bureau. Other situations, including research applications, expert analysis, and model-based adjustments, may have different model-acceptance criteria.

Despite our persistent recommendations for production-ready models, some users are reluctant to change the model that X-12-ARIMA selects. A procedure that performs checks and automatically chooses a new model that fits the criteria will help users who are not confident with their own modeling skills.

Certainly it would be ideal to have experts choose the regARIMA models we use for production work, using all available diagnostics and having full knowledge of each series, but for novices and even for experienced modelers who are short on time and information, automatic modeling procedures are going to continue to be the main source of models used for production work.

We believe the new automatic modeling procedure saves time because it produces generally better models than the previous procedure, and we spend less time testing other models. But when the new procedure selects models that are not production ready, we have to remodel the series.

We compared results of complex models chosen by the automatic modeling procedure to results from using the airline model and simplified models. If the automatically-chosen ARIMA models had no seasonal component, instead of using the true airline model, we used the nonseasonal counterpart, ARIMA (0 1 1). In addition, these models were subject to our unit root checks, so some initial airline models ended up as AR models. However, for sake of simplicity, we continued to refer to this group of models as airline models.

3. Methods

We started with 872 U.S. Census Bureau economic data series, most of which are seasonally adjusted (or adjusted for trading-day effects) on a monthly or quarterly basis. They included data for U.S. imports and exports, construction, manufacturing, retail sales, food stamp participation, and components of quarterly net income after taxes. There were 672 flow series (values that accumulate over time), 17 of which were quarterly series. The remaining 200 series were stock (inventory values measured at a given point in time) series or constructed in such a way as to behave like stock series with regard to trading-day effects. They were a mix of well-behaved and poorly-behaved series, including some series that were not likely to be fit by any model.

We used X-12-ARIMA version 0.3 (build 138, compiled April 26, 2004) for model identification and estimation.

Table 1 shows a typical input specification file for our automatic model identification. The example input specification file is for a monthly flow series. For stock series we tested for stock trading-day effects, and for quarterly series we set period = 4. For 17 series that require prior ratio adjustments, we specified the log transformation instead of using the X-12-ARIMA automatic choice.

This input specification file has X-12-ARIMA perform several automatic adjustment choices:

- Transformation choice (logarithmic or no transformation)
- ARIMA model choice
  - Maximum nonseasonal difference: two
  - Maximum seasonal difference: one
  - Maximum nonseasonal order: three
  - Maximum seasonal order: one
  - No preference for balanced models

- Regression choices
  - Trend constant
  - Trading-day effect
  - Easter effect (test for Easter effect lasting one, eight, or 15 days)
  - Outliers (default critical value)
    - Additive outliers (point outliers)
    - Level shifts
    - Temporary changes

Table 1. Initial Input Specification File

```
series{
  title = 'X0013'
  period = 12
  file = 'X0013.dat'
  format = 'datevalue'
}
transform{ function = auto }
automdl{
  maxdiff = (2,1)
  maxorder = (3,1)
  balanced = no
}
regression{ aictest=( td  Easter ) }
check{print=all}
outlier{types=( all )}
x11 { }
```

We used X-12-Write, a SAS® program that writes and edits X-12-ARIMA input files to create our specification files (Hood 2003). The program runs X-12-ARIMA with automatic options as chosen by the user and creates input specification files specifying the results. Users familiar with X-12-Write may notice that we changed the usual automatic modeling settings for this project.

From the results of X-12-Write, we identified 287 series with models that were not production-ready. Six of those series had no apparent seasonal, trading-day, or Easter effects through any of the X-12-ARIMA runs, so we confined our review to the remaining 281 series.

We are accustomed to seeing models that are not production ready. When changing them manually, we can work back and forth with different regression or ARIMA configurations to choose what is best. However, an automated procedure must be systematic with little backward checking if it is to be fast enough to be useful.

Because changing the model affects the coefficients, we approached coefficient-related changes in steps, not simultaneously:

Step 1. Change mixed models to nonmixed models

Step 2. Change models with problematic coefficients

Our approach to Step 2 was straight-forward. We already had guidelines that we could program fairly closely into X-12-Write. Step 1 was not as simple. We used a system of related models, testing each new combination. We chose model test patterns largely based on experience. Our approach essentially is a hybrid of the procedures found in TRAMO and X-11-ARIMA.

Our approach to the different mixed model patterns is shown in Table 2 (d represents the order of differencing). For models with mixed seasonal components, either $(1\ 0\ 1)_s$ or $(1\ 1\ 1)_s$, we substituted $(0\ 1\ 1)_s$. For models with mixed nonseasonal components, we had up to three test ARIMA models.

Table 2. Substitutions for Nonseasonal Mixed Model Patterns

Mixed Patterns	Substitution Test Patterns
(3 d 3) (3 d 2) (2 d 3)	(0 d 3) (3 d 0)
(3 d 1)	(0 d 3) (3 d 0) (2 d 0)
(2 d 2)	(0 d 2) (2 d 0) (0 d 0)
(2 d 1)	(0 d 2) (2 d 0)
(1 d 3)	(0 d 3)
(1 d 2)	(0 d 2)
(1 d 1)	(0 d 1) (1 d 0) (0 d 0)

Table 3. Second Input Specification File

```
series{
  title = "X0013_2"
  file = 'X0013.dat'
  format = "datevalue"
}
transform{function = log}
arima{ model = (2 2 0)(0 1 1) }
# Auto Choice was (2 2 2)(0 1 1)
regression{ AICtest = ( td  Easter ) }
outlier{ types=( all ) }
check{ print=all }
forecast{ print=none maxlead=30 }
x11{ seasonalma = s3x3 }
```

We ran X-12-ARIMA again, specifying the test models. For comparison purposes, during this step we also created an input file specifying the automatically-identified model. From the results of the automatic identification run, we specified transformation choice, trend constant, the X-11 seasonal moving average, and the number of forecasts. We forecasted half the length of the X-11 seasonal filter (60 months for monthly series with 3x9 filters). We retested for other regression effects because the results, especially outlier results, could be different for different ARIMA models. Table 3 shows an input specification file for this step.

After rerunning X-12-ARIMA and estimating the models for all series, we checked the test model coefficient values for unit root

problems according to the criteria shown in Table 4. We use the following notation: (p d q)(P D Q) is the initial ARIMA model, ϕ_i is the ith nonseasonal AR coefficient of the initial model, θ_i is the ith nonseasonal MA coefficient, Φ_s is the first (only) seasonal AR coefficient, and Θ_s is the first (only) seasonal MA coefficient, and se_{x_n} is the estimated standard error of the model parameter x_n. All coefficients that were within 0.04 of 1.0 were less than 1.96 standard errors from 1.0.

Table 4. Changes for Problematic Coefficients

Coefficient Value	Model Change
Nonseasonal Models	
$\lvert\phi_1 - 1\rvert < 0.04$ and $\lvert\phi_1 - 1\rvert < 1.96\ se_{\phi_1}$	If d = 0 then ({p – 1} 1 0)
or $\sum_{i=1}^{p} \phi_i > 0.96$	If d > 0 (0 d 3)
$\lvert\theta_1 - 1\rvert < 0.04$ and $\lvert\theta_1 - 1\rvert < 1.96\ se_{\theta_1}$	(1 {d – 1} 0) (plus constant for d – 1 = 0)
or $\sum_{i=1}^{q} \theta_i > 0.96$	(no substitution for d = 0)
Seasonal Models	
$\lvert\Phi_s - 1\rvert < 0.04$ and $\lvert\Phi_s - 1\rvert < 1.96\ se_{\Phi_s}$	$(0\ 1\ 1)_s$
$\lvert\Theta_s - 1\rvert < 0.04$ and $\lvert\Theta_s - 1\rvert < 1.96\ se_{\Theta_s}$	If D = 1 then use seasonal dummy regressors (no substitution for D = 0)

For the nonseasonal components, if there was an apparent AR unit root problem, but there was already some order of differencing, we changed to an MA model of the highest order that we were allowing (three).

We considered substituting $(1\ 0\ 0)_s$ + seasonal dummy regressors in place of $(0\ 1\ 1)_s$ where Θ_s was close to one. For some early test runs (with all regressors specified), we checked the significance of Φ_s. We made this substitution for 149 series: 21.5% (32) had at least one test run for which Φ_s was significant (more than 1.96 se_{Φ_s} from zero). Five of those 32 series also had test runs for which it was not significant. (Each series had multiple test runs with different ARIMA models but these counts include only the models with this substitution.)

We also looked at Ljung-Box Q results using the pass/fail criteria we describe below. Including the seasonal AR made a difference for only 16.1% (24) of the series. For 58.3% (14) of those, models with $(1\ 0\ 0)_s$ and seasonal dummies had satisfactory Ljung-Box Q criteria, and models with only seasonal dummies failed. (The estimated Φ_s was significant for half (7) of those series.) For the other 41.7% (10), the situation was reversed, and the model passed only without the seasonal AR. (Only two of those 10 series had significant Φ_s.)

We used seasonal dummies with no seasonal ARMA parameters for our substitutions. But we decided that the program should allow users to choose whether or not to use $(1\ 0\ 0)_s$ with the seasonal dummies.

From the results of the substitution runs, we specified all chosen regressors, specifying up to 20 outliers – if the program selected more than 20 outliers, we did not specify any outliers. We used automatic outlier identification again for all series for our final run, but we set the critical value to 5.5, much higher than for the previous runs. Our previous runs used the default critical value which is derived based on the length of the model span. Table 5 shows an example of a final input specification file with model information specified.

Table 5. Final Test Input Specification File

```
series{
  title = "X0013_2_F"
  savelog = peaks
  file = 'X0013.dat'
  format = "datevalue"
}
transform{ function = log }
arima{ model = (2 2 0)(0 1 1) }
regression{
  variables = ( AO2001.Jul AO2001.Oct )
}
outlier{ critical=5.5 types=( all ) }
check{ print=all }
forecast{ print=none maxlead=30 }
x11{ seasonalma = s3x3 }
history{
  estimates = (fcst sadj sadjchng)
  start = 1999.Jan
}
```

Once we had completed all the substitutions and had run the series with the options specified, we compared models and adjustments using standard diagnostics. We used Ljung-Box Q and the spectrum of the model residuals as qualifying diagnostics. Models that passed our criteria for these were automatically considered better than models that may have had smaller forecast errors or revisions but did not pass the criteria for the first two diagnostics. In addition, we relied on out-of-sample forecast error graphs only for models that were not distinguished using diagnostics 1–4.

1. Ljung-Box Q, goodness-of-fit diagnostics related to the autocorrelation function (Ljung and Box 1978). Models failed if
   a. Lag s Q diagnostic failed, or
   b. More than s lags failed from lag 1 to 2s, or
   c. More than s/2 lags failed from lag 1 to s;
2. Spectrum of the model residuals, a diagnostic of residual seasonal and trading-day effects: models failed if there was a visually significant peak at seasonal frequency 1/s (Soukup and Findley 1999);

3. Average absolute percent error of within-sample forecasts for the last three years of the series, preferring smaller values;
4. Average absolute revisions of the seasonally adjusted series (percentages for multiplicative decompositions), preferring smaller values.
5. Out-of-sample forecast error graphs, preferring the model with generally smaller errors.

We chose not to use likelihood statistics because we did not want diagnostics that we could use for some series and not for others. We often compared models with different orders of differencing and different outlier sets.

We preferred diagnostics that we could incorporate into our program, but we included out-of-sample forecast performance because it is one of the best available model diagnostic tools. This diagnostic consists of a plot of the differences of the accumulating sums of the squared forecast errors of two regARIMA models. We usually plot both the 1-step-ahead and *s*-step-ahead differences. If the differences have a predominantly upward tendency, then the first model has generally larger forecast errors, so we prefer the second model. A predominantly downward tendency indicates a preference for the first model. See Findley et al. (1998) for more information about out-of-sample forecast error graphs.

When modeling a series manually, we would not necessarily use these diagnostics exactly in the way we set them up for this program. For instance, a modeler would likely use autocorrelation and partial autocorrelation function results. But programming a check for those results was too complicated, so we relied on Ljung-Box Q diagnostics.

4. Example, Construction Expenditures Series

An example may clarify the methods we used. For one of the construction expenditure series, X-12-ARIMA chose (2 2 2)(0 1 1) and some outliers for the model. We ran three test runs, testing again each time for regressors:
Model 1. (0 1 1)(0 1 1),
Model 2. (0 2 2)(0 1 1),
Model 3. (2 2 0)(0 1 1), and
Model 4. (0 2 0)(0 1 1).
We also ran X-12-ARIMA with the automatically-chosen model specified.

There were no coefficient problems with the resulting estimates, so we compared the results.

All four of our test models fail the Ljung-Box Q criteria, but each passes our criterion for the spectrum of the model residuals, so we compare average absolute within-sample forecast error and average absolute percent revision.

As shown in Table 6, of all the test runs, Model 1, the airline model, results in the minimum forecast error and the minimum revision, so Model 1 is our choice among the test models.

Table 6. Comparison of Diagnostics Among Models

Test Models	Forecast Error	Revision
1. (0 1 1)(0 1 1)	7.35%	1.465%
2. (0 2 2)(0 1 1)	10.88%	1.942%
3. (2 2 0)(0 1 1)	9.79%	1.965%
4. (0 2 0)(0 1 1)	8.00%	2.212%
Automatic Model		
(2 2 2)(0 1 1)	12.12%	—

We then compare our best test run to the run with the automatically-identified model. The automatically-identified model also failed the Ljung-Box Q criteria but passed the spectrum of model residuals criterion, so we can compare within-sample forecast and revision diagnostics.

Our test run using the airline model has the better within-sample forecast error performance. We cannot compare the revision performance because during the history analysis, the automatically-identified model did not always converge. The history analysis directly imitates production by adding one point at a time, each time reestimating the model. When history estimation fails to converge, we are especially reluctant to use the model for production.

We prefer Model 1 over the automatically-identified model based on these diagnostics, so we do not consider the out-of-sample forecast performance.

5. Results

Of the 872 series, X-12-ARIMA fit 57.5% (501 series) with production-ready models. Another 9.6% (84) resulted in no model, possibly because of convergence problems. This result does not necessarily reflect badly on X-12-ARIMA. Some of the series that we tried to model were so poorly-behaved that we would never try to adjust them. We included all series within our data groups with no attempt to avoid bad series. It turned out that 0.7% (six) of the series had no seasonal, trading-day, or Easter effects in any run, so we excluded them from further tabulations because we would not adjust them in any way.

The remaining 32.2% (281) of the automatically-chosen models were not production-ready: 18.5% (161) were mixed models, and 13.8% (120) were not mixed but had apparent unit-root problems according to our coefficient tests. (Percentages do not add because of rounding.)

Table 7 shows the most frequent nonseasonal mixed-model patterns. The most frequent actual choices were (1 1 1) (30.4%) and (3 1 1) (21.6%).

Table 7. Nonseasonal Mixed-Model Patterns

Pattern	Frequency	Number of Series
(1 d 1)	39.2%	49
(3 d 1)	23.2%	29
(2 d 1)	13.6%	17
Other	24.0%	30
Total	100.0%	125

There were 47 series with mixed seasonal components: 68.1% (32) with $(1\ 0\ 1)_s$ and 31.9% (15) with $(1\ 1\ 1)_s$.

Only 6.8% (11) of the mixed models had mixed nonseasonal and seasonal model components. They are included in the counts shown above. The most frequent doubly-mixed model patterns were (1 d 1)(1 D 1) and (3 d 1)(1 D 1), each chosen for 36.4% (four) of the 11 series.

For the nonmixed models with apparent unit root problems and for 3.7% (six) of the mixed models, we simply made our replacements, so there was just one test model. (These mixed models were all of the form (0 1 1)(1 *d* 1) so that the only test model was the airline model.)

Of the 155 remaining series with model comparisons, 63.9% (99) had an obviously preferred model. These models either (1) passed the Ljung-Box Q and model-residual spectrum diagnostics and the other test models for the series did not (or passed one when all other test models passed neither), or (2) when compared to the other models that passed the qualifying diagnostics (to the greatest degree) had runs that resulted in both the minimum within-sample forecast error and minimum revision. This second situation was the case for our example series of Section 5.

The remaining 36.1% (56) had no test model preference. We created out-of-sample forecast error graphs to compare the models with the highest passing percentage of the qualifying diagnostics. Obviously, this diagnostic check cannot be automated because it requires a choice made according to visual clues. We can automate the creation of the graphs, however, or if we want to eliminate any steps that are not automatic, we must find another way to choose models when the choice is not obvious.

Out-of-sample forecast error graphs helped us choose models for an additional 20% (31) of the 155 series. We were left with 16.1% (25) that still needed a model choice. We decided to use the (possibly modified) airline models for these series.

Next we compared results of the 281 chosen replacement models to the results using the automatically-chosen models. In practice, we would prefer the replacement models over the automatically-chosen models merely to avoid using mixed models or models with possible unit roots in a production setting, but we wanted to see if our changes involved diagnostic penalties.

Again, we used Ljung-Box Q and model residual criteria to compare results. Of the 281 model comparisons, 17.8% (50) had better results using the automatically-chosen model. Of those 50 comparisons, 90.0% (45) favored the automatically-chosen model because it passed the Ljung-Box Q criteria, and the replacement model did not. This result is not surprising: 82% (37) of those 45 models were mixed, and it is reasonable that the more parameterized mixed models would have better goodness-of-fit diagnostics. The qualifying criteria favored the replacement model for 6.4% (18) of the 281 series.

We compared within-sample forecast error and revision of the seasonally adjusted series for the remaining 213 series: 22.1% (47) of the comparisons showed that the automatic model choice was better for both diagnostics, and 38.5% (82) of the comparisons favored the replacement model.

But how different were the diagnostics we compared? To answer this question, we looked at some simple descriptive statistics of the differences for the 213 series. We constructed the differences so that a positive value indicated a preference for the replacement model. Keep in mind that the forecast error diagnostic is the average absolute percent error, and the revision diagnostic is either the average absolute percent difference (for series with multiplicative decompositions) or the average absolute difference (for series with additive decompositions).

Table 8. Descriptive Statistics for the Differences of Average Absolute Percent Within-Sample Forecast Error

		Extreme Values	
Mean	Median	Negative	Positive
1.91	0.22	– 16.6	162.2
		– 9.9	63.4
t =2.19		– 9.2	44.1

Table 8 shows these statistics for the forecast errors and Figure 1 shows a scatter plot of the values. The correlation coefficient was 0.967. We removed the extreme pair (434.8,272.6) from the plot. We can see from Table 8 that the mean, with a t value of 2.19, is significantly different from zero, so the replacement models had better within-sample forecast error on average.

When comparing the revisions, we saw convergence problems. For 1.9% (four) of the 213 series, the revision diagnostic was not available for either the automatic model run or for the replacement model run. For an additional 2.3% (five) of the series, it was not available for the automatic model run but was available for the replacement model run. Certainly for those five series, we prefer the replacement model.

For the remaining 204 series, we separated comparisons by type of decomposition so that we did not compare percentages to values from the original series level. Table 9 shows descriptive statistics for the revision values. Apart from one additive decomposition, the revision values were not especially different. The t values show that the means are not significantly different from zero.

Figure 1. Scatter Plot of the Average Absolute Within-Sample Forecast Error of the Automatic Model Choices and the Replacement Models With Reference Line at Y=X

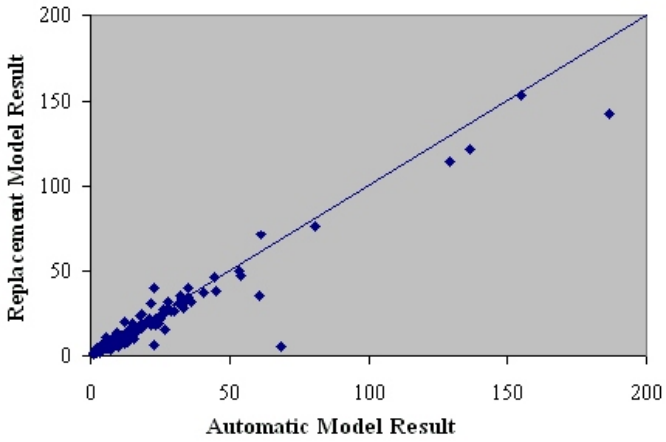

Table 9. Descriptive Statistics for the Differences of Average Absolute Revision of the Seasonally Adjusted Series

		Extreme Values	
Mean	Median	Negative	Positive
Multiplicative Decomposition			
– 0.02	0.00	– 2.35	1.66
		– 0.71	0.55
t = – 0.65		– 0.68	0.34
Additive Decomposition			
0.92	0.02	– 0.37	69.92
		– 0.14	0.54
t = 1.04		– 0.12	0.28

Figure 2. Scatter Plot of the Average Absolute Revision of the Seasonally Adjusted Series Using the Automatic Model Choices and the Replacement Models (Additive Decompositions) With Reference Line at Y=X

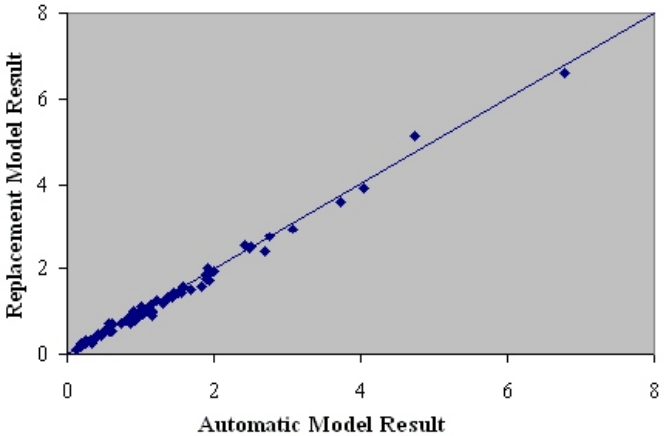

Figure 2 shows a scatter plot of revisions for the 79 series with additive decompositions. We eliminated the extreme pairs (20.455,19.912) and (129.404,59.487) from the plot. The correlation coefficient was 0.984.

Figure 3. Scatter Plot of the Average Absolute Percent Revision of the Seasonally Adjusted Series Using the Automatic Model Choices and the Replacement Models (Multiplicative Decompositions) With Reference Line at Y=X

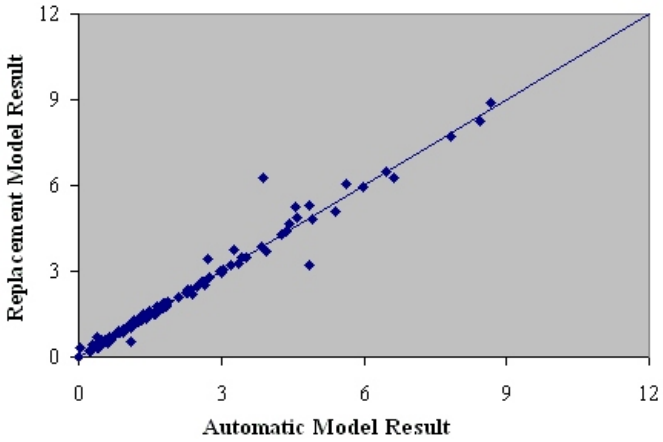

Figure 3 shows a scatter plot of revisions for the 125 series with multiplicative decompositions. The correlation coefficient was 0.986. The overall correlation coefficient for the 204 comparison series was 0.970.

6. Conclusion

We concluded that this automatic model-simplification procedure can help us quickly choose production-ready models whose results compare favorably to results from the automatically-identified models. To fully automate the current procedure, we need to make slight changes. We anticipate that staff at the Census Bureau will be able to use this program.

On the other hand, a new X-12-ARIMA automatic-modeling option is available: setting mixed = no will prevent the procedure from considering mixed models. We are investigating the new option, but meanwhile, we have continued developing the model-simplification program. Our program results are subject to comparisons to results from this new option.

7. Future Study

If our methodology proves valuable even in light of new options of X-12-ARIMA, we would like to further our work with model simplification. Some important issues that remain include checking replacement models for coefficient significance and testing the final model for unit root problems. Also we would like to simulate complicated ARIMA processes to check our results for cases in which we know the true underlying model.

Also we would like to improve the efficiency of the simplification program to reduce the time to run and compare series.

We received a suggestion (B. Monsell, personal communication, September 8, 2004) that different automatic modeling options could

reduce the number of mixed models that the program identifies. We would like to investigate the suggestion further.

Finally, we have many series that we modeled with the possibly-modified airline model. We chose these models for 15.7% (44) of the series based on the diagnostic results and applied them for an additional 8.9% (25) of the series for which the diagnostics did not lead to a particular model. In addition, we would like to look at the results more closely to see how the results from the airline model differ from the results of our other replacement models.

Acknowledgments

We would like to thank Brian C. Monsell and Michael Z. Shimberg of the U.S. Census Bureau for their valuable comments and suggestions. Additional thanks to Brian for X-12-ARIMA programming support. We also appreciated Ayonda Dent's help in interpreting out-of-sample forecast error graph results.

References

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

Dagum, E. B. (1980), "The X-11-ARIMA Seasonal Adjustment Method," Statistics Canada.

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 0.3," 2001 Proceedings of the American Statistical Association, Business and Economic Section [CD-ROM], Alexandria, VA: American Statistical Association, paper 00100.pdf.

Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C. , and Chen, B.- C. (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 Maravall, A. (1997), "Program TRAMO and SEATS: Instructions for the User, Beta Version," Banco de España.

Hood, C. C. (2003), "X-12-Write: A SAS® Program to Write Input Files, Version 1.1, SAS Version 8," U.S. Census Bureau, U.S. Department of Commerce.

Hood, C. C., and Findley, D. F. (1999), "An Evaluation of TRAMO/SEATS and Comparison with X-12-ARIMA," Proceedings of the American Statistical Association, Business and Economic Statistics Section, 150-155.

Ljung, G. M. and Box, G. E. P. (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., Young, A. H., and Musgrave, J. C. (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 Findley, D. F. (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 (2004), *X-12-ARIMA Reference Manual, Version 0.3 (Beta)*, U.S. Census Bureau, U.S. Department of Commerce.