Determining Seasonality: A Comparison of Diagnostics From X-12-ARIMA

Demetra P. Lytras, Roxanne M. Feldpausch, and William R. Bell
U.S. Census Bureau

Keywords: Seasonal adjustment, Diagnostics

1. Introduction

Before a series is seasonally adjusted, it should be shown that the series is seasonal. When using X-12-ARIMA for seasonal adjustment, two diagnostics commonly used to determine seasonality are M7, a diagnostic developed at Statistics Canada for X-11-ARIMA, and the F-test for seasonality assuming stability, referred to as the D8 F-test for stable seasonality; however, the statistical properties of these are not well understood. In this paper, we examine properties of these statistics and compare them to those of two other diagnostics available in the program: the presence of seasonal peaks in the spectrum of the differenced original series, and a model-based chi-squared test of fixed seasonal effects. We also examine a modification of the chi-squared test to give the corresponding model-based F-test (F^M). We use simulated nonseasonal series to determine and compare the significance levels of these diagnostics and simulated seasonal series to assess their power.

In Section 2, we discuss the various methods for determining seasonality. We discuss the methodology used to find the significance levels and present these results in Section 3. In Sections 4 and 5 we explain the methodology used to evaluate the tests' power and present these results. We summarize our results in Section 6.

2. Background

At the U.S. Census Bureau, we use the D8 F-test for stable seasonality, along with M7 and the spectrum of the differenced original series, to determine whether or not a series is seasonal. We investigate the properties of these diagnostics, along with the chi-squared and F^M tests for fixed seasonal effects, using various simulated nonseasonal and seasonal regARIMA models. These models have the general form

$$(1-B)^d(1-B^s)^D\phi(B)\Phi(B^s)(Y_t - \beta'X_t) = \theta(B)\Theta(B^s)\alpha_t$$

where B is the backshift operator, such that $BY_t=Y_{t-1}$, $\phi(B)$, $\Phi(B^s)$, $\theta(B)$, and $\Theta(B^s)$ are nonseasonal and seasonal autoregressive (AR) and moving-average (MA) polynomials of orders p, P, q, and Q, respectively, $\beta'X_t$ are fixed regression effects, Y_t is the observed time series, and α_t is white noise with mean zero and variance σ^2. A particular model is often identified by the orders of its differencing and AR and MA polynomials as (p d q)(P D Q). For example, Box and Jenkins's (1976) classic airline model is an ARIMA model of the form (0 1 1)(0 1 1), and so it has one nonseasonal and one seasonal difference, $\theta(B)=(1-\theta B)$, and $\Theta(B)=(1-\Theta B^s)$. In this paper we assume the time series are monthly.

2.1 D8 F-test for Stable Seasonality

The D8 F-test checks for the equality of the monthly means; that is, it tests the hypothesis

H_0: $m_1=m_2=\ldots=m_{12}$
H_1: $m_p \neq m_q$ for at least one pair (p,q)

where $m_1,\ldots,m_{12}$ are the monthly means of the seasonal-irregular (SI) component (the detrended series) found in table D8. It assumes that the SI values are independently distributed as normal with means m_i and common standard deviation σ. However, while this could be true conceptually for the underlying true SI ratios, the *estimates* of the SI ratios are actually dependent and heteroscedastic, which affects the behavior of the resulting F-statistic. The traditionally attempted solution to this problem is to not use a critical value from the F-distribution, but to instead use a cut-off value of 7, with values greater than 7 indicating that the series is seasonal (McDonald-Johnson et al. 2006). For more information on the D8 F, see Ladiray and Quenneville (2001).

2.2 M7

M7 is one of the monitoring and quality assessment statistics developed by Statistics Canada in the 1970s (Lothian and Morry 1978). M7 values less than one are interpreted as indicating that the series has identifiable seasonality, while values greater than one are interpreted as indicating that either the series is not seasonal or the seasonality cannot be identified by the X11 algorithm. M7 is calculated using the D8 F-test for stable seasonality

and the D8 F-test for moving seasonality, which tests whether the seasonality changes over the years, using the equation

$$\text{M7}=\sqrt{\frac{1}{2}\left(\frac{7}{F_s}+\frac{3F_m}{F_s}\right)}$$

where F_s is the D8 F-statistic for stable seasonality and F_m is the D8 F-statistic for moving seasonality (Ladiray and Quenneville 2001).

2.3 Spectrum of the Differenced Series

X-12-ARIMA estimates the spectrum of the differenced original series as follows. Let $w_t = Y_t - Y_{t-1}$ be the first differenced series. Then the value of the spectrum at the frequency λ is

$$\hat{s}(\lambda) = 10 \log_{10}\left\{\frac{\hat{\sigma}_{30}^2}{2\pi \left|1-\sum_{j=1}^{30}\hat{\phi}_j e^{ij2\pi\lambda}\right|^2}\right\}$$

for $0 \le \lambda \le 0.5$, where the coefficient estimates $\hat{\phi}_j$ are those from the linear regression of $w_t - \overline{w}$ on $w_{t-j} - \overline{w}$, $1 \le j \le 30$, and $\hat{\sigma}_{30}^2$ is the sample variance of the resulting regression residuals. X-12-ARIMA calculates values of $\hat{s}(\lambda)$ at 61 frequencies and graphs the results in the output file using lineprinter plots. The seasonal frequencies (1/12, 2/12,…,6/12) are marked on the graph (U.S. Census Bureau 2007). A six "star" peak at one of the seasonal frequencies is considered "visually significant," with one star corresponding to 1/52nd of the range between the maximum and minimum spectral values (Soukup and Findley 1999). By default, X-12-ARIMA uses the last eight years of the first differenced original series in its calculations of the spectrum. We only considered frequencies 1/12,…,4/12 (McDonald-Johnson et al. 2006).

2.4 Model-Based Chi-squared and F-tests for Fixed Seasonal Effects

When fixed seasonal effects are specified, X-12-ARIMA fits a regARIMA model with the following 11 variables included in the vector X_t:

$$M_{1,t} = \begin{cases} 1 & \textit{in January} \\ -1 & \textit{in December}, \ldots, \\ 0 & \textit{otherwise} \end{cases} \quad M_{11,t} = \begin{cases} 1 & \textit{in November} \\ -1 & \textit{in December} \\ 0 & \textit{otherwise} \end{cases}.$$

The model fitting provides estimates of the regression parameters $\hat{\beta}$ and corresponding t-statistics that can be used to determine if the individual parameter estimates are significantly different from zero, and also a chi-squared test statistic for testing if the parameters are collectively all zero (no fixed seasonal effects). The chi-squared test statistic is

$$\hat{\chi}^2 = \hat{\beta}'[Var(\hat{\beta})^{-1}]\hat{\beta}.$$

The statistic is compared to critical values from the χ^2-distribution with 11 degrees of freedom. When calculating this statistic X-12-ARIMA uses its estimate of the innovation variance in calculating $Var(\hat{\beta})$. In referring this statistic to the χ^2 distribution we effectively assume that the estimated innovation variance equals the true innovation variance. While this is true asymptotically (under suitable assumptions), for finite samples the estimation of the innovation variance affects the behavior of the test statistics. We can, however, correct the chi-squared test to account for the estimation of the innovation variance by using the corresponding test statistic F^M, calculated as follows:

$$F^M = \frac{\hat{\chi}^2}{11} \times \frac{n-d-k}{n-d}$$

Here $\hat{\chi}^2$ is the chi-squared statistic from above, n is the number of observations in the series, d is the degree of differencing, and k is the total number of elements estimated in β. (If the only regression effects are the fixed seasonal effects, then k = 11; however, if there are other regression effects in the model then k > 11.) F^M follows an $F_{11,n-d-k}$ distribution.

3. Nonseasonal Series—Calculating Significance Levels

3.1 Methods

We simulated nonseasonal series to check the significance levels of the four diagnostics. We simulated 10,000 series of length 20 years from each of the following seven models:

ARIMA (0 1 0)
ARIMA (0 1 1), with θ=0.3, 0.5, and 0.8
ARIMA (1 1 0), with ϕ=0.3, 0.5, and 0.8

The series were each run in X-12-ARIMA Version 0.3 Build 174 using a regARIMA model with the correct ARIMA model and with fixed seasonal effects. We ran the series both with and without the correct ARMA parameters specified to check whether the performance of the diagnostics deteriorates when the parameters are estimated. In each run, X-12-ARIMA estimated the innovation variance, as there is no way to specify it in the program. We ran X-12-ARIMA with the default settings

in the `x11` spec and two years of forecasts, and both with the spectrum calculated using the default 8 years of data and using all 20 years of data.

3.2 Results

The significance levels of the five diagnostics for series from each of the seven ARIMA models, as found when X-12-ARIMA was run with the correct model, estimated parameters, and fixed seasonal effects, are shown in Table 1. Note that the standard error of the significance levels would be 0.002 for p=0.05, 0.003 for p=0.1, and as high as 0.005, when p=0.5. (From considerations of the binomial distribution the standard error is $[p(1-p)/10{,}000]^{.5}$.)

As Table 1 shows, the significance levels of the M7 and D8 F-statistics and the spectrum peaks diagnostic vary greatly depending on the ARIMA model from which the series was simulated. These three diagnostics all find fewer (0 1 1) series to be seasonal than (1 1 0) series. The significance levels of all three decrease as the θ parameter increases in (0 1 1) series, while in the (1 1 0) series the significance levels of M7 and the D8 F increase as ϕ does and the significance levels of the spectrum peaks decrease as ϕ increases. For all models, fewer seasonal series are identified when the spectrum is calculated over twenty years than over the default eight years.

The results of the model-based chi-squared test of the fixed seasonal effects are more consistent: for each model, 7.5%−8% of the simulated nonseasonal series have significant fixed seasonal effects detected when tested at the 0.05 level. With the correction for the estimation of the innovation variance (F^M), this proportion is closer to 5%. (As a check we ran the simulated series through X-12-ARIMA using the correct model and the correct parameters, and found that F^M matched its stated significance levels, as it should have.)

Table 1: Significance Levels of the Seasonality Diagnostics

Model	M7	D8F	Spectrum Peaks 8 years	Spectrum Peaks All Data	SR chi-squared p=0.05	F^M p=0.05
(0 1 1)						
θ= 0.3	0.0094	0.0008	0.1893	0.1424	0.0789	0.0536
0.5	0.0016	0.0000	0.1437	0.0802	0.0758	0.0538
0.8	0.0000	0.0000	0.0892	0.0298	0.0772	0.0527
(1 1 0)						
ϕ= 0.3	0.0816	0.0628	0.2132	0.1711	0.0769	0.0548
0.5	0.1063	0.1118	0.1791	0.1113	0.0786	0.0541
0.8	0.1087	0.1676	0.1041	0.0304	0.0777	0.0534
(0 1 0)	0.0312	0.0116	0.2353	0.2262	0.0752	0.0505

Because of the large range of significance values found for the various models and parameters, size-adjusted critical values were found for the M7, D8 F, and spectrum peak diagnostics for each of the models, using the series run with the correct model and parameters. These are the values such that when the diagnostics are applied to (nonseasonal) series from the given models, they will exceed, or in the case of M7 be less than, the critical values in Table 2 5% of the time. Remember that, normally, series are considered seasonal if M7 is less than 1, D8 F is greater than 7, and there is a seasonal peak in the spectrum of the differenced original series greater than 6 (stars).

Table 2: Critical Values for a Significance Level of 0.05

Model	Parameter	M7	D8 F	Spectrum Peak Height (in stars) 8 years	Spectrum Peak Height (in stars) All Data
(0 1 1)	0.3	1.223	3.337	12.7	10.1
	0.5	1.348	2.712	10.8	7.5
	0.8	1.513	2.145	7.9	5.0
(1 1 0)	0.3	0.918	7.542	13.7	11.0
	0.5	0.858	9.736	11.6	8.7
	0.8	0.847	12.154	8.1	5.1
(0 1 0)		1.065	4.854	13.9	13.7

4. Seasonal Series—Calculating Power

4.1 Methods

We explored how well these four diagnostics correctly identified seasonality using simulated series with fixed seasonal effects, and also with series simulated from the airline model. Tables 3 and 4 below report the results.

The series with fixed seasonal effects were generated by adding a fixed set of seasonal factors to 1,000 simulated nonseasonal series. This was done for all 36 combinations from each of the six (0 1 1) and (1 1 0) models of Table 1 with six sets of seasonal factors. The six sets of seasonal factors were generated from the two sets of factors shown in Figure 1, multiplying them by one of three constants so the maximum magnitude of the seasonal factors was either less than, close to, or greater than the innovation standard deviation (1.0). The six sets of fixed seasonal factors thus had:

Set 1a: Values ranging from –0.42 to 0.70
Set 1b: Values ranging from −0.57 to 0.96
Set 1c: Values ranging from –1.31 to 2.18
Set 2a: Values ranging from –0.65 to 0.64
Set 2b: Values ranging from –0.98 to 0.96
Set 2c: Values ranging from –1.73 to 1.70

The series were run in X-12-ARIMA with the correct ARIMA model, estimated parameters, and including fixed seasonal effects in the model. The power of the seasonal

diagnostics was found when they were compared to the size adjusted critical values from Table 2.

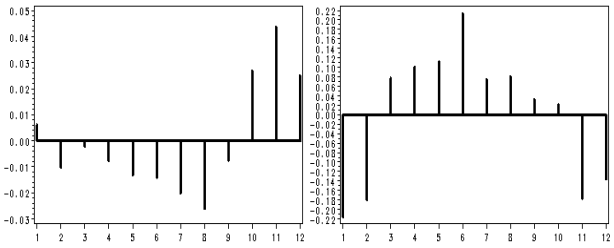

Figure 1: Seasonal factor sets 1 and 2

We also simulated 1,000 series from each of several airline models with various starting values, series lengths, innovation variances, and MA parameters. Series were simulated for each combination of the following:

Seasonal $\Theta = 0.6$ and 0.9
Nonseasonal $\theta = 0.3$ and 0.8
Series length of 10 years and 20 years
Small, medium, and large innovation variances
Starting values of all zeroes, as well as two additional sets of starting values given in Figure 2.

We ran the series in X-12-ARIMA with the regARIMA model identified as (0 1 1) + fixed seasonal effects + trend constant, allowing X-12-ARIMA to estimate all model parameters.

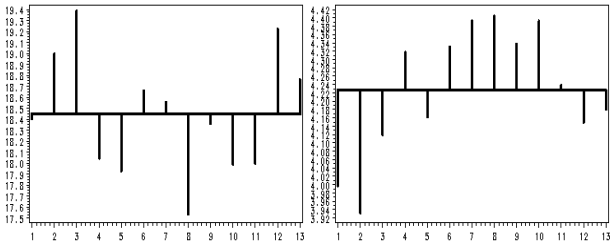

Figure 2: Starting values used for airline series

4.2 Results

The power of all four diagnostics varies when fixed seasonal effects of varying patterns and magnitudes are added to the previously generated nonseasonal series. Table 3 shows the proportions of series found to be seasonal when the fixed seasonal effects sets are added to the various simulated nonseasonal series. For the M7, D8 F, and spectrum diagnostics we used the size-adjusted critical values shown in Table 2. For the model-based test we used F^M to correct the small size distortions of the chi-squared statistic. Note that the standard errors for these values would be 0.009 when p=0.9, and as large as 0.016 when p=0.5.

The F^M test for the significance of the fixed seasonal effects is, overall, the most powerful of the diagnostics. At the 0.05 level the F^M test identifies as seasonal all the series in sets 1c and 2c (where the magnitude of the fixed seasonal effects are large compared to the innovation standard deviation) and almost all the series in sets 1b and 2b (where the values of the fixed seasonal effects range to about one, which is the innovation standard deviation). In sets 1a and 2a, the magnitude of the fixed seasonal effects is always less than the innovation standard deviation of the series. The power of the F^M test remains near one for set 2a, but for set 1a the power is between 0.61 and 0.70 for the (1 1 0) series at the 0.05 level.

Table 3: Power of Diagnostics When Using the Size-Adjusted Critical Values

					Spectrum Peaks		F^M,
Set	Model	ϕ,θ	M7	D8 F	8 years	All Data	p=0.05
1a	(0 1 1)	0.3	0.856	0.872	0.202	0.378	0.843
		0.5	0.948	0.948	0.227	0.478	0.923
		0.8	0.959	0.965	0.258	0.399	0.957
	(1 1 0)	0.3	0.334	0.367	0.109	0.215	0.613
		0.5	0.240	0.246	0.122	0.210	0.586
		0.8	0.163	0.156	0.123	0.216	0.699
1b	(0 1 1)	0.3	0.990	0.995	0.363	0.737	0.992
		0.5	1.000	1.000	0.436	0.822	0.999
		0.8	1.000	1.000	0.494	0.849	1.000
	(1 1 0)	0.3	0.649	0.677	0.188	0.413	0.933
		0.5	0.415	0.436	0.204	0.391	0.914
		0.8	0.296	0.300	0.217	0.456	0.961
1c	(0 1 1)	0.3	1.000	1.000	0.949	1.000	1.000
		0.5	1.000	1.000	0.982	1.000	1.000
		0.8	1.000	1.000	0.993	1.000	1.000
	(1 1 0)	0.3	1.000	1.000	0.704	0.990	1.000
		0.5	1.000	1.000	0.758	0.997	1.000
		0.8	0.968	0.984	0.863	1.000	1.000
2a	(0 1 1)	0.3	0.964	0.973	0.236	0.466	0.964
		0.5	0.993	0.995	0.309	0.647	0.985
		0.8	1.000	1.000	0.434	0.739	1.000
	(1 1 0)	0.3	0.516	0.544	0.182	0.324	0.978
		0.5	0.339	0.367	0.185	0.394	0.989
		0.8	0.186	0.198	0.238	0.535	1.000
2b	(0 1 1)	0.3	1.000	1.000	0.518	0.874	1.000
		0.5	1.000	1.000	0.621	0.962	1.000
		0.8	1.000	1.000	0.794	0.995	1.000
	(1 1 0)	0.3	0.884	0.907	0.355	0.689	1.000
		0.5	0.702	0.749	0.444	0.818	1.000
		0.8	0.510	0.533	0.563	0.944	1.000
2c	(0 1 1)	0.3	1.000	1.000	0.926	1.000	1.000
		0.5	1.000	1.000	0.970	1.000	1.000
		0.8	1.000	1.000	0.993	1.000	1.000
	(1 1 0)	0.3	1.000	1.000	0.784	0.996	1.000
		0.5	0.998	1.000	0.907	1.000	1.000
		0.8	0.963	0.971	0.958	1.000	1.000

Table 4: Power of Seasonal Diagnostics Using the Airline Series

				Series length = 10 years				Series length = 20 years			
Starting Values	σ^2	Θ	θ	M7	D8 F	Spectrum Peaks	F^M, p=0.05	M7	D8 F	Spectrum Peaks	F^M, p=0.05
Zero	1	0.6	0.3	0.989	0.972	0.979	1.000	0.999	0.999	0.995	1.000
Zero	1	0.6	0.8	0.972	0.942	0.948	0.999	1.000	0.998	0.992	1.000
Zero	1	0.9	0.3	0.974	0.923	0.920	1.000	0.997	0.990	0.954	1.000
Zero	1	0.9	0.8	0.969	0.901	0.897	1.000	0.997	0.989	0.902	0.999
Zero	0.1	0.6	0.3	0.977	0.964	0.967	0.999	1.000	0.999	0.996	1.000
Zero	0.1	0.6	0.8	0.981	0.953	0.959	1.000	1.000	0.999	0.995	1.000
Zero	0.1	0.9	0.3	0.981	0.939	0.924	0.998	0.998	0.996	0.952	1.000
Zero	0.1	0.9	0.8	0.960	0.892	0.900	1.000	0.997	0.991	0.921	1.000
Set 1	1	0.6	0.3	0.994	0.991	0.987	1.000	0.999	1.000	0.998	1.000
Set 1	1	0.6	0.8	0.997	0.990	0.983	1.000	1.000	0.999	0.994	1.000
Set 1	1	0.9	0.3	0.996	0.985	0.980	1.000	1.000	0.999	0.983	1.000
Set 1	1	0.9	0.8	0.990	0.977	0.968	1.000	1.000	0.998	0.951	1.000
Set 1	0.136	0.6	0.3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Set 1	0.136	0.6	0.8	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
Set 1	0.136	0.9	0.3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Set 1	0.136	0.9	0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Set 2	1	0.6	0.3	0.984	0.962	0.987	0.999	0.999	1.000	0.995	1.000
Set 2	1	0.6	0.8	0.978	0.960	0.956	0.999	1.000	1.000	0.993	1.000
Set 2	1	0.9	0.3	0.984	0.945	0.941	0.999	0.997	0.994	0.946	1.000
Set 2	1	0.9	0.8	0.971	0.904	0.910	1.000	1.000	0.997	0.921	1.000
Set 2	0.015	0.6	0.3	1.000	1.000	0.996	1.000	1.000	1.000	0.998	1.000
Set 2	0.015	0.6	0.8	1.000	0.999	0.999	1.000	1.000	1.000	0.997	1.000
Set 2	0.015	0.9	0.3	1.000	0.998	0.996	1.000	1.000	1.000	0.994	1.000
Set 2	0.015	0.9	0.8	1.000	0.999	0.997	1.000	1.000	1.000	0.996	1.000

The M7 and the D8 F show a great deal more variation of their power across the different models, and in some cases the power is rather low. All the diagnostics performed well for all models when the fixed seasonal effects were large compared to the innovation standard deviation (sets 1c and 2c). For the other sets, M7 and D8 F were powerful for (0 1 1) series, particularly as the MA parameter grew larger. However, M7 and D8 F performed poorly for the (1 1 0) series, particularly as the AR parameter became larger; for both sets 1a and 2a, these two diagnostics found less than 20% of the series with ϕ=0.8 to be seasonal.

The spectrum was consistently the least powerful of the diagnostics, particularly when it used the default setting of 8 years. Like the other diagnostics, it had more trouble identifying (1 1 0) series as seasonal. However, it was the only diagnostic to also have extremely low power in identifying (0 1 1) series as seasonal.

Table 4 shows the proportion of airline series detected as seasonal for each combination of starting values, innovation variance, model length, and model parameters. The F^M test found nearly every series seasonal, while the M7, D8 F, and spectral peaks varied some in how well they detected seasonality amongst the models, with the spectral peaks generally doing the worst job. (In Table 4 the spectrum diagnostic used just 8 years of data.) However, even these three diagnostics indicated that at least about 90% of the series were seasonal in all cases.

For each model, M7, D8 F, and spectral peaks had higher power when used on longer series. In the case of the series with nonzero starting values, seasonality was detected by M7, D8 F, and the spectrum for almost all series when the innovation variance σ^2 was set to the sample variance of the starting values, and the power decreased slightly when the innovation variance increased. The value of Θ also appeared to have a small

effect on the number of series found to be seasonal; slightly more seasonal series were identified when Θ was 0.6 than 0.9, when all other values are held constant. In many cases these differences were statistically significant.

5. Incorrect Model Choice

The results in Sections 3 and 4 assumed the correct regARIMA models were used. In practice, we cannot know the correct model. To test how these diagnostics perform when an incorrect model is specified, we ran the nonseasonal series and the series with fixed seasonal effects through X-12-ARIMA with an incorrect ARIMA model specified. To specify an incorrect model we ran all the (1 1 0) series as (0 1 1), and we ran all the (0 1 1) series as (1 1 0).

The significance levels, size-adjusted critical values, and power of the M7, D8 F, and spectrum diagnostics were identical or very close to their values when the program was run with the correct model. As these diagnostics are not strongly dependent on the model specification, the model only being used for forecast and backcast extension, this was an expected result. There were some differences in the significance levels and powers of the F^M test, however. Table 5 shows the significance levels of F^M when using correct and incorrect models. For the incorrect models Table 6 gives the size adjusted critical values for a 0.05 significance level. Table 7 compares the powers of the diagnostics under the correct versus incorrect models.

Table 5: Significance Levels With the Correct and the Incorrect Models Specified (p=0.05)

Real Model	ϕ,θ	F^M, Correct Model	F^M, Incorrect Model
(0 1 1)	0.3	0.0536	0.0618
	0.5	0.0538	0.0722
	0.8	0.0527	0.0944
(1 1 0)	0.3	0.0548	0.0492
	0.5	0.0541	0.0395
	0.8	0.0534	0.0125

Table 6: Critical Value for the Model-Based F-test When the Wrong Model is Specified (Actual Critical Value is 1.831)

Real Model	ϕ,θ	F^M, Incorrect Model
(0 1 1)	0.3	1.899
	0.5	1.969
	0.8	2.085
(1 1 0)	0.3	1.827
	0.5	1.760
	0.8	1.356

Table 7: Power With the Correct and the Incorrect Models Specified (p=0.05)

Set	Model	ϕ,θ	F^M, Correct Model	F^M, Incorrect Model
1a	(0 1 1)	0.3	0.843	0.827
		0.5	0.923	0.867
		0.8	0.957	0.779
	(1 1 0)	0.3	0.613	0.589
		0.5	0.586	0.512
		0.8	0.699	0.409
1b	(0 1 1)	0.3	0.992	0.992
		0.5	0.999	0.994
		0.8	1.000	0.985
	(1 1 0)	0.3	0.933	0.922
		0.5	0.914	0.863
		0.8	0.961	0.780
1c	(0 1 1)	0.3	1.000	1.000
		0.5	1.000	1.000
		0.8	1.000	1.000
	(1 1 0)	0.3	1.000	1.000
		0.5	1.000	1.000
		0.8	1.000	1.000
2a	(0 1 1)	0.3	0.964	0.950
		0.5	0.985	0.950
		0.8	1.000	0.921
	(1 1 0)	0.3	0.978	0.978
		0.5	0.989	0.988
		0.8	1.000	0.996
2b	(0 1 1)	0.3	1.000	1.000
		0.5	1.000	1.000
		0.8	1.000	1.000
	(1 1 0)	0.3	1.000	1.000
		0.5	1.000	1.000
		0.8	1.000	1.000
2c	(0 1 1)	0.3	1.000	1.000
		0.5	1.000	1.000
		0.8	1.000	1.000
	(1 1 0)	0.3	1.000	1.000
		0.5	1.000	1.000
		0.8	1.000	1.000

One of the benefits of using the F^M test rather than one of the other diagnostics when the correct model was specified was that the power was consistent and very close to the nominal 5% across the different models. This is not always the case when an incorrect model is used. When (0 1 1) series were misspecified as (1 1 0) series, the significance level was larger than the expected 0.05 for all parameters, and as θ increased, so did the significance level. The opposite occurred when (1 1 0) series were specified as (0 1 1); in all cases, the significance level was smaller than the expected 0.05 for all parameters, and this significance level decreased as ϕ increased.

The actual critical value for a 0.05 level of an $F_{11,228}$ distribution, 1.831, is used to find the power of the F-test when the correct model is given. These results are in the first column of Table 7; the second column shows the power when the incorrect model is specified, using the size-adjusted critical values given in Table 6. When the actual parameter is 0.3, the power is either identical or slightly less with the incorrect model. The largest difference is 0.024. The difference grows as the parameter becomes larger, with the largest difference in (1 1 0) models with a parameter of 0.8 in Set 1a; the power drops from 0.699 to 0.409 when the incorrect model is used.

While the size and power of the F^M test can be affected by specifying an incorrect model, note that these effects are quite mild for ϕ or θ equal to 0.3, are still not that serious for ϕ or θ equal to 0.5, and really only become substantial for ϕ or θ equal to 0.8. Moving through these parameter values from 0.3 to 0.5 to 0.8 takes us to models that are more and more incorrect, in the sense that for low values of |ϕ| the (0 1 1) model can approximate the (1 1 0) model quite well, but not for large values of of |ϕ|, and similarly for small versus large values of |θ| in the (0 1 1) model.

But at the same time when |ϕ| is large we are unlikely to misspecify a (1 1 0) series as (0 1 1), and similarly when |θ| is large we are unlikely to mis-specify a (0 1 1) series as (1 1 0). This is true whether the model specification comes from an automatic criterion, such as AIC, or from examination of sample autocorrelation and partial autocorrelation functions. In general, small model misspecifications are more likely than large ones, and the above results suggest that small model misspecifications may lead to only mild deterioration in the performance of F^M. A more relevant evaluation may be to use a model selection criterion like AIC to pick models for the simulated series, rather than deliberately specifying an incorrect (or a correct) model. We intend to pursue this in future research.

6. Summary

The diagnostics M7, D8 F, and the spectrum peaks of the differenced original series have significance levels that vary greatly amongst the different models. This indicates that the traditional cutoff values would need to be modified based on the model and the model parameters in order to provide consistent results. This is not a practical solution, since in practice the true model and parameter values are not known, so the actual correct critical values would not be known. As an approximation one could consider taking the estimated model as truth, determining the appropriate critical value for this model, and using that value. However, this would be clumsy (perhaps requiring an additional simulation), and error in estimating the model parameters would affect the determination of the critical values, and thus the properties of the resulting diagnostics.

The model-based chi-squared test has a consistent significance level, but is somewhat oversized. Fortunately, using the F^M test substantially corrects the size distortion, and the power of this test is higher than or consistent with those of the other diagnostics. Specifying an incorrect model has some effect on the size and power of the F^M test, but these effects appear mild for mild deviations from the correct model, and major deviations from the correct model are less likely to occur and more readily detected and corrected. Thus, overall the F^M test appears to perform the best among the available seasonality diagnostics. The F^M test also has the advantage (relative to M7, D8 F, and the spectrum diagnostic) that appropriate critical values for the test are readily available from the standard table of the F-distribution.

References

Box, G.E.P. and Jenkins, G.M. (1976), *Time Series Analysis: Forecasting and Control*, Holden Day: San Francisco.

Ladiray, D. and Quenneville, B. (2001), *Seasonal Adjustment with the X-11 Method.* Springer-Verlag: New York, NY.

Lothian, J. and Morry, M. (1978), "A Set of Quality Control Statistics for the X-11-ARIMA Seasonal Adjustment Method," Research Paper, Seasonal Adjustment and Time Series Staff, Statistics Canada.

McDonald-Johnson, K., Monsell, B., Fescina, R., Feldpausch, R., Hood, C., and Wroblewski, M. (2006), *Seasonal Adjustment Diagnostics: Census Bureau Guideline*, Washington, DC: 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," *1999 Proceedings of the American Statistical Association*, Business and Economic Statistics Section, Alexandria, VA: American Statistical Association: pp. 144-149.

U.S. Census Bureau (2007), *X-12-ARIMA Reference Manual, Version 0.3*, Washington, DC: U.S. Census Bureau, U.S. Department of Commerce. http://www.census.gov/ts/x12a/v03/x12adocV03.pdf