RESEARCH REPORT SERIES
(Statistics #2016-02)

Examining Diagnostics for Trading-Day Effects from X-13ARIMA-SEATS

Osbert Pang
Brian Monsell

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: March 11, 2016

Examining Diagnostics for Trading-Day Effects from X-13ARIMA-SEATS

Osbert Pang and Brian Monsell

Abstract

This study examines some diagnostics available in X-13ARIMA-SEATS for detecting a trading-day effect. The diagnostics of interest are a χ^2-test, an F-test, and the spectrum peaks at the two tested trading-day frequencies (under both the default last 8 years of data as used by the program and the full data). Sets of seasonal series without trading-day effects are simulated initially to measure the false detection rate and to estimate the appropriate $\alpha = 0.05$ critical value. Sets of seasonal series with trading-day effects are subsequently simulated to assess the power of those diagnostics.

1 Introduction

Trading-day effects can occur in a series when activity varies depending on the day of the week. Since months are not uniform in length, aggregating this daily activity to a monthly total may result in variation that is attributable to the composition of days in a month. For example, retail series might see slightly higher sales for months that contain more Saturdays and Sundays. Bell and Hillmer (1983) suggested a regression-type model to account for trading-day and other holiday effects; combined with an autoregressive integrated moving average (ARIMA) model, the overall model could handle both time series behavior and calendar variation. X-13ARIMA-SEATS contains multiple diagnostics that can be used for detecting the presence of a trading-day effect. These range from the statistically-based χ^2- and F-tests on a trading-day regressor to more informal tests of "visual significance" with respect to various spectrum estimators (see Cleveland and Devlin (1980) for an early discussion of spectrum analysis for detecting trading-day and other calendar effects).

Soukup and Findley (1999) previously conducted a simulation study to examine the spectrum diagnostics available in X-12-ARIMA for detecting a trading-day effect. Using a base set of 42 real (monthly) economic series that had trading-day effects present, they ran these series through X-12-ARIMA without estimating a trading-day effect. From examining the spectrum estimators of some of the resulting output series, they reported the detection rate as the percentage of series for which

X-12-ARIMA found a "visually significant" peak at a particular trading-day frequency. To obtain simulated series for establishing false detection rates, they took a multiplicative decomposition of each series into a trend, seasonal, and irregular component; since no trading-day effect was estimated in the original series, all trading-day variation should be contained within the irregular component. Replicates of each series were produced by bootstrapping the respective irregular component. These replicates thus preserved the trend and seasonal aspects of the original series, but not the observed trading-day pattern. The false detection rate was determined by running these replicate series through X-12-ARIMA with the same procedure used to establish the detection rate. The spectrum estimators used in the simulation study of Soukup and Findley (1999) appear to have been computed using the default last 8 years of data, and the study concluded that the presence of a "visually significant" peak in either the spectrum estimator of the differenced seasonally adjusted series or that of the irregular component was the preferred diagnostic for determining a trading-day effect. A "6-star" peak was determined to satisfy the condition of "visual significance," and their study found that this criteria had a false detection rate between 10 and 15 percent.

Lytras et al. (2007) performed a different type of simulation study to assess the diagnostics for detecting seasonality in X-12-ARIMA. Using some of the methodology described there, Ladiray (2012) conducted a similar study on detection of trading-day effects. The simulation study of Ladiray (2012) used a set of 22 nonseasonal and seasonal ARIMA models, with the idea that each model aligned with a particular spectrum estimator when detecting a trading-day effect using "visual significance." This visual test was conducted for both the autoregressive spectrum estimator that was the default option in X-12-ARIMA and a periodogram estimate. In addition to examining the test of "visual significance," the statistically-based diagnostics available in X-12-ARIMA were also considered. The series simulated were of variable length, and both monthly and quarterly series were used. The study suggested that the periodogram estimate may be preferable to the autoregressive spectrum estimator as far as the visual tests were concerned, but that the regARIMA F-test performed best for detecting a trading-day effect.

The simulation study described in this report is more similar in construction to that of Ladiray (2012) than to that of Soukup and Findley (1999), but restricts itself to monthly series of length 20 years. The length itself is somewhat arbitrary. The other simulation studies appear to have examined only the spectrum estimator produced using the default last 96 observations (i.e., the last 8 years for a monthly series). There is also some difference from Ladiray (2012) in that four spectrum peak diagnostics associated with the autoregressive spectrum estimator are evaluated, although there is no tweaking of the width of the band used for determining spectral peaks as there was in Ladiray (2012).

Like Ladiray (2012), we examine the regARIMA χ^2- and F-tests for a fixed trading-day effect as well as the various spectrum peak diagnostics reported by X-13ARIMA-SEATS (at the two tested trading-day frequencies using both the X-13ARIMA-SEATS default setting and the full series). We start by laying out the framework for the simulations, namely the models and parameter sets used. Using a simulated set of series with no trading-day effect, we study the sizes of the various diagnostic

tests. Adding trading-day effects to those previously simulated series allows for an examination of the power of those same diagnostic tests.

Section 2 provides the framework for the simulation study. Section 3 uses series with no trading-day effect to look at the false detection rate for the diagnostics considered. Section 4 uses the set of $\alpha = 0.05$ size-adjusted critical values derived from the results of Section 3 to simulate the power for these diagnostics. Section 5 re-examines the false detection rate and power for some other autoregressive spectrum estimators produced by X-13ARIMA-SEATS as a comparison to the previous results obtained for the spectrum estimator of the regARIMA model residuals. Section 6 summarizes the results.

2 Methodology

For the simulation study, monthly series of length 20 years are generated for 24 different seasonal ARIMA models. Ladiray (2012) used a set of nonseasonal and seasonal models as reflective of the cases encountered when testing spectrum peak diagnostics in X-13ARIMA-SEATS: the spectrum associated with white noise was analogous to the spectra of model residuals and the irregular component, the spectrum associated with nonseasonal models was analogous to that of seasonally adjusted series, and the spectrum associated with seasonal models was analogous to that of the original series. The following models are used in this simulation:

- $(0\ 1\ 1)_{12}$, with $\theta_{12} = 0.3$, 0.5, and 0.8
- $(0\ 1\ 0)(0\ 1\ 1)_{12}$, with $\theta_{12} = 0.3$, 0.5, and 0.8
- $(0\ 1\ 1)(0\ 1\ 1)_{12}$, with $\theta_1 = 0.3$, 0.5, and 0.8, and $\theta_{12} = 0.3$, 0.5, and 0.8
- $(1\ 1\ 0)(0\ 1\ 1)_{12}$, with $\phi_1 = 0.3$, 0.5, and 0.8, and $\theta_{12} = 0.3$, 0.5, and 0.8

Since these models all have the general form $(p\ d\ q)(0\ 1\ 1)_{12}$, each can be represented as

$$(1 - \phi_1 B)^p (1 - B)^d (1 - B^{12}) y_t = (1 - \theta_1 B)^q (1 - \theta_{12} B^{12}) \xi_t,$$

with p, d, and q equal to zero or one as appropriate, and where ξ_t is some zero-mean white noise process (for convenience, a normal distribution is typically chosen); this simulation assumes a normal distribution with mean 0 and variance 1 for ξ_t. Solving for y_t provides a recursive formula that can be used to simulate series for each model.

For the χ^2- and F-tests, the simulated series are modeled in X-13ARIMA-SEATS using the correctly specified regARIMA model, along with regressors for a constant term and a trading-day effect; the estimated trading-day effect will be tested at an $\alpha = 0.05$ significance level. Note also that X-13ARIMA-SEATS provides a χ^2-test for both a trading-day regressor and a combined trading-day and leap year regressor. The results of the two χ^2-tests should not be appreciably different, but only the one for the trading-day regressor will be considered. This is because the

F-statistic is a linear transformation of this χ^2 statistic ($F = \frac{\chi^2}{6} \times \kappa$, where $\kappa = \frac{227}{219}$ for the $(0\ 1\ 1)_{12}$ models and $\frac{228}{220}$ for the others). The false detection rate of the two statistical tests is approximately the proportion of simulated series for which the p-value of the corresponding test statistic computed by X-13ARIMA-SEATS is less than 0.05.

The diagnostics for spectrum peaks use a non-statistical test based on the autoregressive spectral density estimator. In X-13ARIMA-SEATS the spectrum is plotted at frequencies λ_k of the form $k/120$, $0 \leq k \leq 61$. The software also computes the spectrum at the trading-day frequencies (0.348 and 0.432) along with the frequencies 1/120 before and after these trading-day frequencies (i.e., $\{0.339\overline{6}, 0.348, 0.356\overline{3}\}$ and $\{0.423\overline{6}, 0.432, 0.440\overline{3}\}$). A peak at a trading-day or seasonal frequency is declared "visually significant" if it is both greater than the median of plotted values and larger than its neighbors by at least 6/52 times the range of plotted values (a "star" is equivalent to 1/52 times the range, so a "visually significant" peak is 6 "stars" high). More details can be found in Section 6.1 of the X-13ARIMA-SEATS reference manual (U.S. Census Bureau (2015)).

The values of the spectrum at the two trading-day frequencies can be extracted from the diagnostic summary file (.udg file). For evaluating the spectrum peaks, the simulated series are modeled in X-13ARIMA-SEATS using the correctly specified regARIMA model and a constant term, but no trading-day regressor. This is done twice – once using all 20 years for the spectrum and once using just the last 8 years. The peak strength will be saved as is for any peak that is both greater than the median and larger than its neighbors, but will be set to 0 otherwise. Under this framework, the false detection rate is the proportion of simulated series for which the value at either of the two trading-day frequencies exceeds 6. Soukup and Findley (1999) observed that a threshold of 6 yields a false detection rate of approximately 0.1 when evaluated at just the 0.348 trading-day frequency.

3 False Detection Rates

A simulation run of 10,000 monthly series for each seasonal ARIMA model was performed. The spectrum peak diagnostic examined was the spectrum of the model residuals. Table 1 has the false detection rate for the two statistically-based diagnostics (χ^2 and F), as well as for the peak diagnostic for the spectrum of the model residuals for each of the seasonal ARIMA models.

The results shown are more or less as expected. The F-test comes closest to achieving the desired size, although the false detection rates observed would still be considered significantly different for more than half of the models. Ladiray (2012) found smaller false detection rates for the F-test than were obtained here (and were lower than the assumed 0.05 significance level). While his numbers were obtained by averaging results over models with different parameter values and series of different lengths, those factors would not account for this observed difference. We were unable to replicate Ladiray's numbers, as our results suggest there is instead a small upward size bias. The χ^2 performs slightly worse in this respect than the F-test, which is attributable to the fact that the F-test accounts for estimation of the variance (and the χ^2 does not). The "6-star" rule for spectrum peaks of model residuals has higher false detection rates than either of the statistically-based tests,

				χ^2	F	Residual Spectrum	
Model	ϕ_1	θ_1	θ_{12}	p = 0.05	p = 0.05	8 years	All 20
$(0\ 1\ 1)_{12}$			0.3	0.0624	0.0481	0.1409	0.1395
			0.5	0.0647	0.0506	0.1365	0.1414
			0.8	0.0656	0.0526	0.1323	0.1492
$(0\ 1\ 0)(0\ 1\ 1)_{12}$			0.3	0.0671	0.0519	0.1372	0.1363
			0.5	0.0703	0.0554	0.1325	0.1406
			0.8	0.0637	0.0513	0.1351	0.1523
$(0\ 1\ 1)(0\ 1\ 1)_{12}$		0.3	0.3	0.0690	0.0546	0.1391	0.1396
			0.5	0.0744	0.0581	0.1343	0.1395
			0.8	0.0749	0.0591	0.1357	0.1536
		0.5	0.3	0.0707	0.0561	0.1397	0.1391
			0.5	0.0684	0.0550	0.1370	0.1439
			0.8	0.0672	0.0538	0.1353	0.1548
		0.8	0.3	0.0656	0.0524	0.1400	0.1348
			0.5	0.0648	0.0518	0.1404	0.1465
			0.8	0.0645	0.0497	0.1363	0.1492
$(1\ 1\ 0)(0\ 1\ 1)_{12}$	0.3		0.3	0.0749	0.0597	0.1373	0.1363
			0.5	0.0741	0.0594	0.1335	0.1358
			0.8	0.0734	0.0592	0.1309	0.1604
	0.5		0.3	0.0716	0.0573	0.1344	0.1368
			0.5	0.0698	0.0570	0.1398	0.1482
			0.8	0.0693	0.0551	0.1327	0.1512
	0.8		0.3	0.0679	0.0538	0.1358	0.1444
			0.5	0.0623	0.0477	0.1396	0.1492
			0.8	0.0593	0.0449	0.1365	0.1516

Table 1: False detection rates for trading-day diagnostics.

and the rates we find are larger than the 0.1 rate observed by Soukup and Findley (1999). This is likely a by-product of our testing two frequencies. Compared to the false detection rates observed by Ladiray (2012) for a (0 0 0) model (white noise), however, we find rates that are noticeably lower. We also observe slightly higher rates associated with the spectrum constructed using the full data and for the series with a seasonal θ_{12} of 0.8. Overall, the numbers are reasonably stable, hovering around the 0.13 to 0.16 range. This stability suggests that one can achieve a desired false detection rate using the spectrum of the model residuals by changing the threshold rule from 6 stars to some appropriate value.

As a quick check, the simulation is repeated, but with fixed parameter values for the models rather than estimated values. Under these conditions, the F-test should detect a trading-day effect 5% of the time, with some slight deviation that can be attributable to simulation error. Looking at each of the 24 models individually, there are 2 for which the F-test has a significance level that would be significantly different from 0.05, but averaging over the 24 models, the overall significance level is 0.0493, which is reasonably close to the desired 0.05.

4 Simulated Power

In this study, the reference distribution for the χ^2-test is a χ^2_6 for all of the models. For the F-test, the appropriate reference distribution is an $F_{6,220}$ for all but the $(0\ 1\ 1)_{12}$ and an $F_{6,219}$ for the $(0\ 1\ 1)_{12}$ models. Using an $\alpha = 0.05$ significance level, the resulting critical values are approximately

						Peak Heights (stars)	
Model	ϕ_1	θ_1	θ_{12}	χ^2	F	8 years	All 20
$(0\ 1\ 1)_{12}$			0.3	13.1551	2.1156	11.6	11.3
			0.5	13.3548	2.1477	11.7	11.6
			0.8	13.4268	2.1593	11.5	11.6
$(0\ 1\ 0)(0\ 1\ 1)_{12}$			0.3	13.4338	2.1601	12.2	11.7
			0.5	13.6349	2.1924	11.6	11.5
			0.8	13.4225	2.1582	11.8	12.3
$(0\ 1\ 1)(0\ 1\ 1)_{12}$		0.3	0.3	13.6154	2.1893	11.9	12.0
			0.5	13.6799	2.1996	12.0	11.6
			0.8	13.8720	2.2305	11.8	12.1
		0.5	0.3	13.6108	2.1885	12.5	11.5
			0.5	13.6083	2.1881	12.0	11.5
			0.8	13.5230	2.1744	11.9	12.1
		0.8	0.3	13.4222	2.1582	12.4	11.8
			0.5	13.3755	2.1507	12.3	12.2
			0.8	13.2840	2.1360	11.6	12.0
$(1\ 1\ 0)(0\ 1\ 1)_{12}$	0.3		0.3	13.8603	2.2286	12.1	11.4
			0.5	13.8757	2.2311	11.7	11.3
			0.8	13.8628	2.2290	11.7	12.4
	0.5		0.3	13.7434	2.2098	12.3	11.7
			0.5	13.7247	2.2068	12.3	11.9
			0.8	13.6674	2.1976	11.7	12.1
	0.8		0.3	13.4843	2.1682	12.0	11.9
			0.5	13.1742	2.1183	11.9	11.8
			0.8	13.0055	2.0912	12.1	12.3

Table 2: Simulated critical values for a significance level of 0.05.

12.6 for the χ^2-test and roughly 2.14 for the F-test. However, for power calculations, we find the size-adjusted critical value by determining the 95th percentile of the observed test statistics from the above simulations. Table 2 gives the size-adjusted critical values for each of the diagnostics described in the previous section. We can see that the ranges are actually fairly small for all of the diagnostics, as all of the critical values are between 13 and 13.9 for the χ^2-test, between 2.09 and 2.23 for the F-test, and between 11.3 and 12.5 for the peak diagnostic applied to the residual spectra (the residual spectrum constructed using all 20 years of data tends to have slightly smaller critical values than the one constructed using just the last 8 years). Note that the size-adjusted χ^2 critical value for each model in our simulation exceeds the actual $\alpha = 0.05$ critical value for a χ^2_6 distribution, and the same is true in all but a few cases with respect to the size-adjusted F critical value. As mentioned previously, the F-statistic can be obtained directly from the χ^2 statistic (and vice versa); as a result, the power calculations for both diagnostics should be identical when size-adjusted critical values are used. Because of this, since the F-test comes closer to achieving the nominal size, the power results will be reported for the F-test, but not for the χ^2-test.

With the simulated series from the previous section, we can construct series with fixed trading-day effects. Following the notation of Ladiray (2012), if a simulated series is labeled as ARIMA_t and the trading-day effect as TD_t, then a series with a trading-day effect is just $X_t = \text{ARIMA}_t + \delta\text{TD}_t$, where δ represents a scale factor for the magnitude of the trading-day effect. We use 3 different trading-day patterns that have mean 0 and variance 1; these are plotted in Figure 1 and were chosen to approximate the three different trading-day patterns described by Ladiray (2012). The

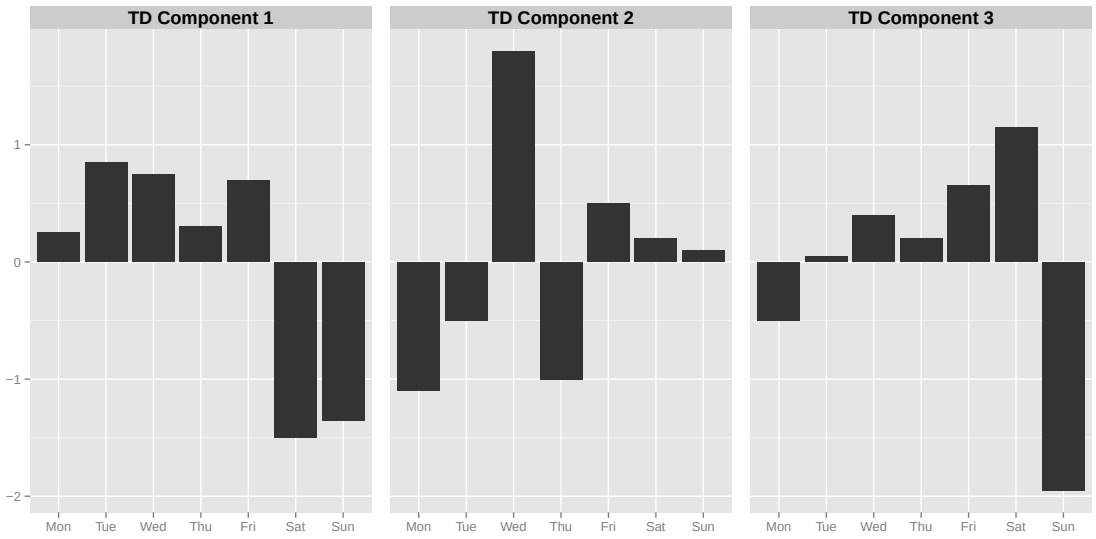

Figure 1: Trading-day components used for simulation.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
TD Component 1	0.25	0.85	0.75	0.30	0.70	−1.50	−1.35
TD Component 2	−1.10	−0.50	1.80	−1.00	0.50	0.20	0.10
TD Component 3	−0.50	0.05	0.40	0.20	0.65	1.15	−1.95

Table 3: The daily effects associated with the three trading-day patterns.

exact numbers used for the approximation are provided in Table 3.

In constructing the various series X_t for this simulation, δ values of 0.1, 0.3, and 0.5 are used. Each series is then modeled in X-13ARIMA-SEATS using the correctly specified ARIMA model with the parameters estimated. As before, a trading-day regressor will be included for the F-test, but will be excluded for spectrum peaks. The power of each diagnostic is estimated by finding the proportion of series that exceed the corresponding size-adjusted critical values. We display simulated power results for $\delta = 0.1$ (Table 4) and $\delta = 0.3$ (Table 5), averaging the simulated power observed over the three trading-day components. The power results for $\delta = 0.5$ will be omitted due to the fact that the power is very close to 1 in all but a few cases.

From the two tables, it is apparent that the peak diagnostic for the spectrum of the model residuals has lower power than the F-test does. This is in line with what was seen in Ladiray (2012). However, we observed that the F-test has higher power (averaging over the three trading-day patterns) for $\delta = 0.1$ and $\delta = 0.3$ than was found by Ladiray (2012). The two studies do not use identical sets of models, so the averages are not necessarily comparable, but the difference is noticeable nevertheless. It is also unsurprising that the disparity in power between the spectrum peaks diagnostic and the statistical test is greatest when the trading-day effect is weak. Ladiray (2012) does not appear to observe a noticeable difference in power between series of varying lengths for the visual tests, which might suggest that the spectrum estimator used the same length (perhaps the default last 8 years of data) regardless of series length. The results obtained here indicate that

					Residual Spectrum	
Model	ϕ_1	θ_1	θ_{12}	F	8 years	All 20
$(0\ 1\ 1)_{12}$			0.3	0.3876	0.0953	0.1202
			0.5	0.3656	0.0935	0.1249
			0.8	0.3148	0.0847	0.1158
$(0\ 1\ 0)(0\ 1\ 1)_{12}$			0.3	0.8126	0.1777	0.2896
			0.5	0.7813	0.1906	0.3128
			0.8	0.7296	0.1659	0.2654
$(0\ 1\ 1)(0\ 1\ 1)_{12}$		0.3	0.3	0.6741	0.1457	0.2106
			0.5	0.6542	0.1433	0.2299
			0.8	0.5789	0.1315	0.2072
		0.5	0.3	0.5779	0.1157	0.1826
			0.5	0.5580	0.1251	0.1913
			0.8	0.4993	0.1130	0.1681
		0.8	0.3	0.4425	0.0989	0.1382
			0.5	0.4338	0.0982	0.1392
			0.8	0.3800	0.0964	0.1332
$(1\ 1\ 0)(0\ 1\ 1)_{12}$	0.3		0.3	0.8996	0.2370	0.4075
			0.5	0.8781	0.2541	0.4283
			0.8	0.8302	0.2174	0.3659
	0.5		0.3	0.9437	0.2699	0.4664
			0.5	0.9278	0.2845	0.4848
			0.8	0.8894	0.2633	0.4410
	0.8		0.3	0.9842	0.3569	0.5517
			0.5	0.9761	0.3680	0.5891
			0.8	0.9566	0.3229	0.5351

Table 4: Average power of trading-day diagnostics when δ is 0.1.

					Residual Spectrum	
Model	ϕ_1	θ_1	θ_{12}	F	8 years	All 20
$(0\ 1\ 1)_{12}$			0.3	0.9967	0.3623	0.5957
			0.5	0.9910	0.3676	0.6052
			0.8	0.9755	0.3319	0.5691
$(0\ 1\ 0)(0\ 1\ 1)_{12}$			0.3	1.0000	0.5526	0.8280
			0.5	1.0000	0.5959	0.8417
			0.8	1.0000	0.5435	0.8087
$(0\ 1\ 1)(0\ 1\ 1)_{12}$		0.3	0.3	1.0000	0.5399	0.7750
			0.5	1.0000	0.5449	0.7903
			0.8	0.9996	0.5044	0.7611
		0.5	0.3	0.9999	0.4686	0.7328
			0.5	0.9998	0.4948	0.7484
			0.8	0.9985	0.4446	0.7119
		0.8	0.3	0.9994	0.3963	0.6476
			0.5	0.9970	0.4051	0.6537
			0.8	0.9897	0.3816	0.6282
$(1\ 1\ 0)(0\ 1\ 1)_{12}$	0.3		0.3	1.0000	0.6836	0.9046
			0.5	1.0000	0.7032	0.9118
			0.8	1.0000	0.6484	0.8856
	0.5		0.3	1.0000	0.7164	0.9261
			0.5	1.0000	0.7221	0.9343
			0.8	1.0000	0.6990	0.9210
	0.8		0.3	1.0000	0.7490	0.9525
			0.5	1.0000	0.7529	0.9599
			0.8	1.0000	0.7198	0.9477

Table 5: Average power of trading-day diagnostics when δ is 0.3.

the power is notably higher when the spectrum is estimated using the full series as opposed to just the last 8 years.

It can also be seen from the Tables 4 and 5 that the simulated power varies across ARIMA models: the diagnostics perform better for detecting a trading-day effect with a $(1\ 1\ 0)(0\ 1\ 1)_{12}$ seasonal ARIMA model than they do for the other models considered here, with the worst performance for the $(0\ 1\ 1)_{12}$ models. The size of the model parameters also influences the power of the various diagnostics. The power generally seems to decrease when seasonal moving average parameter θ_{12} is increased from 0.5 to 0.8. In addition, there appears to be a decrease in power as the nonseasonal moving average parameter θ_1 increases for the $(0\ 1\ 1)(0\ 1\ 1)_{12}$ model and, conversely, an increase in the power is observed as the nonseasonal autoregressive parameter ϕ_1 increases for the $(1\ 1\ 0)(0\ 1\ 1)_{12}$ models. The power scales fairly well as the magnitude of the trading-day effect (as represented by δ) increases. With a δ of 0.5, even the worst-performing diagnostic shown in the tables above – the spectrum peaks diagnostic using just the last 8 years of data for a $(0\ 1\ 1)_{12}$ model – would correctly detect a trading-day effect more than half of the time.

5 Residual Spectrum versus Others

The spectrum peak diagnostic we initially considered was of the regARIMA model residuals when no trading-day regressor was included in the model. X-13ARIMA-SEATS also estimates the spectrum for three other series: the differenced original series, the seasonally adjusted series, and the irregular series. In the interest of comparing the peak diagnostic for the various spectrum estimators, we conduct a second simulation run for these other spectrum peak diagnostics. To provide a consistent base, these other three spectrum diagnostics will also be considered using a regARIMA model with no trading-day regressor. In order to obtain the spectra of the seasonally adjusted series and irregular series, a default seasonal adjustment is performed using the `x11` spec. Table 6 provides the false detection rates using the "6-star" rule for each of these three diagnostics.

The first feature that stands out is that the peak diagnostic applied to the spectrum of the differenced original is highly unlikely to produce a spurious detection. This result is understandable, as the two trading-day frequencies that are tested in X-13ARIMA-SEATS happen to lie close to seasonal frequencies. For the three sets of spectrum peak diagnostics shown in the table, it is also apparent that the false detection rates using the spectrum from the full data are almost universally lower than those using the spectrum from just the last 8 years of data; this was not necessarily the case with the residual spectrum. As was noted with the residual spectrum, it appears that the model has an effect on the detection properties of the diagnostics, as the false detection rates appear to be lower for the $(1\ 1\ 0)(0\ 1\ 1)_{12}$ models compared to the others.

Comparing the numbers observed in the table to those obtained previously for the residual spectrum (see Table 1), it can also be observed that the peak diagnostics for the spectra of the irregular and seasonally adjusted series yield false detection rates that are less stable than those for the residual spectrum, which were found to fall consistently in the 0.13 to 0.16 range. Although

				Last 8 Years			All 20 Years		
Model	ϕ_1	θ_1	θ_{12}	Original	Irregular	SeasAdj	Original	Irregular	SeasAdj
$(0\ 1\ 1)_{12}$			0.3	0.0000	0.1576	0.1080	0.0000	0.1172	0.0544
			0.5	0.0004	0.1292	0.0856	0.0000	0.0852	0.0280
			0.8	0.0016	0.1288	0.0792	0.0000	0.0692	0.0264
$(0\ 1\ 0)(0\ 1\ 1)_{12}$			0.3	0.0000	0.1088	0.1108	0.0000	0.0504	0.0292
			0.5	0.0008	0.1016	0.1192	0.0000	0.0452	0.0416
			0.8	0.0024	0.1176	0.1384	0.0016	0.0516	0.0656
$(0\ 1\ 1)(0\ 1\ 1)_{12}$		0.3	0.3	0.0000	0.1012	0.1176	0.0000	0.0636	0.0524
			0.5	0.0000	0.1232	0.1556	0.0000	0.0620	0.0684
			0.8	0.0028	0.1156	0.1384	0.0012	0.0476	0.0632
		0.5	0.3	0.0004	0.1216	0.1416	0.0000	0.0628	0.0632
			0.5	0.0000	0.1220	0.1452	0.0000	0.0728	0.0876
			0.8	0.0028	0.1020	0.1224	0.0004	0.0512	0.0560
		0.8	0.3	0.0000	0.1280	0.1200	0.0000	0.0872	0.0812
			0.5	0.0004	0.1300	0.1216	0.0000	0.0716	0.0624
			0.8	0.0036	0.1180	0.0988	0.0000	0.0588	0.0460
$(1\ 1\ 0)(0\ 1\ 1)_{12}$	0.3		0.3	0.0004	0.1076	0.0896	0.0000	0.0524	0.0212
			0.5	0.0000	0.0868	0.0816	0.0000	0.0380	0.0200
			0.8	0.0028	0.1096	0.1080	0.0004	0.0580	0.0508
	0.5		0.3	0.0000	0.1056	0.0696	0.0000	0.0564	0.0196
			0.5	0.0000	0.0948	0.0732	0.0000	0.0432	0.0148
			0.8	0.0016	0.0960	0.0756	0.0004	0.0472	0.0288
	0.8		0.3	0.0000	0.1000	0.0372	0.0000	0.0688	0.0040
			0.5	0.0004	0.0904	0.0384	0.0000	0.0428	0.0036
			0.8	0.0016	0.0976	0.0528	0.0000	0.0428	0.0096

Table 6: False detection rates for other spectrum peak trading-day diagnostics.

they are fairly close, the false detection rates for the spectrum peak diagnostic applied to the seasonally adjusted series are slightly lower than those for the spectrum peak diagnostic applied to the irregular series. Soukup and Findley (1999) observed that the spectra of the irregular and seasonally adjusted series, evaluated at the 0.348 trading-day frequency, have false detection rates close to that of the residual spectrum, but this does not appear to be the case in our numbers; this could be a result of our choice to evaluate at both trading-day frequencies. In addition, the false detection rates we obtain here are notably lower than those observed by Ladiray (2012). For example, comparing the spectrum of the irregular series to that of a white noise model, or comparing the spectrum of the seasonally adjusted series to those of nonseasonal models, the (averaged) false detection rates in Ladiray (2012) are higher than what we observed.

Table 7 displays the corresponding size-adjusted (for $\alpha = 0.05$) critical values for the three spectrum peak diagnostics here. While it had been previously observed that the spectrum of the differenced original was highly unlikely to produce a false detection, this spectrum may not be particularly informative, as any positive value for the peak diagnostic would be treated as evidence for a trading-day effect. The spectra for the irregular and seasonally adjusted series may prove more useful, but there is more variation in the critical values for these when compared to the critical values of the model residuals. Whereas the size-adjusted critical values for the spectrum peaks of model residuals all fell in a range between 11.3 (from the 20-year spectrum) and 12.5 (from the 8-year spectrum), the choice of model and span for the spectrum has a more noticeable effect on the magnitude of the critical value for the peak diagnostic in the spectra of both the irregular

				Last 8 Years			All 20 Years		
Model	ϕ_1	θ_1	θ_{12}	Original	Irregular	SeasAdj	Original	Irregular	SeasAdj
$(0\ 1\ 1)_{12}$			0.3	0.0	11.7	8.5	0.0	10.1	6.2
			0.5	0.0	10.2	7.8	0.0	7.8	5.0
			0.8	0.0	10.3	7.4	0.0	7.4	4.4
$(0\ 1\ 0)(0\ 1\ 1)_{12}$			0.3	0.0	9.9	10.6	0.0	6.0	4.4
			0.5	0.0	9.3	10.7	0.0	5.7	5.3
			0.8	0.0	10.2	12.5	0.0	6.1	7.6
$(0\ 1\ 1)(0\ 1\ 1)_{12}$		0.3	0.3	0.0	9.2	11.2	0.0	6.7	6.1
			0.5	0.0	10.4	12.9	0.0	6.9	7.5
			0.8	0.0	9.7	12.2	0.0	5.8	6.8
		0.5	0.3	0.0	10.1	12.3	0.0	6.9	6.8
			0.5	0.0	10.2	12.4	0.0	7.6	8.7
			0.8	0.0	9.6	10.8	0.0	6.0	6.3
		0.8	0.3	0.0	10.5	10.4	0.0	8.1	7.7
			0.5	0.0	10.8	10.0	0.0	7.2	6.9
			0.8	0.0	10.2	8.8	0.0	6.6	5.7
$(1\ 1\ 0)(0\ 1\ 1)_{12}$	0.3		0.3	0.0	9.8	8.7	0.0	6.2	3.4
			0.5	0.0	8.6	7.8	0.0	4.8	3.0
			0.8	0.0	9.6	10.5	0.0	6.3	6.0
	0.5		0.3	0.0	8.9	7.0	0.0	6.6	3.4
			0.5	0.0	9.3	7.7	0.0	5.4	2.6
			0.8	0.0	9.0	8.2	0.0	5.6	4.1
	0.8		0.3	0.0	9.4	5.3	0.0	7.0	2.7
			0.5	0.0	9.0	5.3	0.0	5.4	2.2
			0.8	0.0	9.3	6.1	0.0	5.5	2.6

Table 7: Simulated spectrum peak critical values for a significance level of 0.05.

and seasonally adjusted series. It would appear, then, that the peak diagnostic for the residual spectrum should be favored over the other three spectrum peak diagnostics in X-13ARIMA-SEATS when it comes to detecting the presence of a trading-day effect.

A final check is to compare the power of the various spectrum peak diagnostics. Table 8 displays the power, averaged over both trading-day components and values of δ, for the various spectra estimated using the last 8 years of data. Table 9 provides the analogue for those same spectra estimated using all 20 years of data. Note that the power results for $\delta = 0.5$ are included in the averaging performed for both tables, hence the numbers obtained for the peak diagnostic of the residual spectrum will differ from the average of the corresponding numbers from Tables 4 and 5.

Previously, it was observed that the spectrum peak diagnostic for the differenced original rarely produced a false detection. In the simulations conducted, the spectrum peak diagnostic for the differenced original only exceeds a power of 0.5 in one case, which occurs when the model parameters for a $(1\ 1\ 0)(0\ 1\ 1)_{12}$ model are $\phi_1 = 0.8$ and $\theta_{12} = 0.8$. These problems indicate that the peak diagnostic for the spectrum of the differenced original series should be avoided as a diagnostic for detecting trading-day effects.

For the other three spectrum peak diagnostics, the power is generally 0.1 to 0.2 higher when the spectrum is estimated using the full series as opposed to just the last 8 years. The spectrum of the irregular series and that of the seasonally adjusted series appear to yield similar results for power, although neither fares as well as the spectrum for the model residuals; the advantage of the

Model	ϕ_1	θ_1	θ_{12}	Original	Irregular	SeasAdj	Residual
$(0\ 1\ 1)_{12}$			0.3	0.0267	0.2882	0.2910	0.3385
			0.5	0.0533	0.3124	0.2964	0.3377
			0.8	0.2117	0.2825	0.2859	0.3120
$(0\ 1\ 0)(0\ 1\ 1)_{12}$			0.3	0.0820	0.4096	0.4220	0.4539
			0.5	0.1510	0.4345	0.4340	0.4914
			0.8	0.3713	0.3813	0.3558	0.4592
$(0\ 1\ 1)(0\ 1\ 1)_{12}$		0.3	0.3	0.0589	0.4140	0.3878	0.4669
			0.5	0.1117	0.3840	0.3339	0.4656
			0.8	0.3137	0.3726	0.3157	0.4379
		0.5	0.3	0.0553	0.3749	0.3116	0.4110
			0.5	0.0905	0.3706	0.3013	0.4288
			0.8	0.2828	0.3564	0.3189	0.3932
		0.8	0.3	0.0373	0.3349	0.2921	0.3572
			0.5	0.0692	0.3144	0.2885	0.3647
			0.8	0.2317	0.3036	0.2928	0.3510
$(1\ 1\ 0)(0\ 1\ 1)_{12}$	0.3		0.3	0.1141	0.4165	0.4548	0.5813
			0.5	0.1842	0.4612	0.5000	0.5954
			0.8	0.4283	0.4260	0.4380	0.5550
	0.5		0.3	0.1292	0.4403	0.4730	0.6096
			0.5	0.2174	0.4336	0.4676	0.6157
			0.8	0.4652	0.4671	0.4972	0.5996
	0.8		0.3	0.1643	0.4350	0.4684	0.6518
			0.5	0.2597	0.4604	0.4929	0.6555
			0.8	0.5021	0.4790	0.5127	0.6247

Table 8: Average power for all spectrum peak diagnostics, using just the last 8 years.

residual spectrum varies depending on the model. Previously, it had been observed that power for the residual spectrum had a tendency to decrease for large θ_{12} – this feature occurs less consistently for the spectra of the irregular and seasonally adjusted series. Lastly, it is worth noting that the magnitude of the nonseasonal parameters for the underlying series (θ_1 for the $(0\ 1\ 1)(0\ 1\ 1)_{12}$ models and ϕ_1 for the $(1\ 1\ 0)(0\ 1\ 1)_{12}$ models) also seem to affect the power of the peak diagnostics: the average power tends to be higher for smaller θ_1 or for larger ϕ_1. Overall, though, the comparisons suggest that the peak diagnostic for the spectrum of model residuals is the superior option among these four spectrum peak diagnostics in X-13ARIMA-SEATS.

6 Conclusions

Our simulation study examines some of the diagnostic tests available in X-13ARIMA-SEATS for detecting a trading-day effect. The diagnostics considered were regARIMA χ^2- and F-tests, as well as the spectrum peak diagnostics applied to four of the spectrum estimators produced by X-13ARIMA-SEATS. As was the case in Ladiray (2012), our results indicate that the model-based F-test works best in that it comes closest to achieving the nominal size and has the highest power. Unlike Ladiray (2012), however, our results place the size of the F-test slightly above the $\alpha = 0.05$ level used in this study and not below. Given that the χ^2-tests in X-13ARIMA-SEATS do not improve upon the F-test in either false detection rate or power, they may be redundant.

The four spectrum peak diagnostics examined in this study were the model residuals, the differenced original series, the seasonally adjusted series, and the irregular component. The procedure

Model	ϕ_1	θ_1	θ_{12}	Original	Irregular	SeasAdj	Residual
$(0\ 1\ 1)_{12}$			0.3	0.0175	0.4524	0.4644	0.5120
			0.5	0.0394	0.4722	0.4684	0.5174
			0.8	0.1976	0.4405	0.4479	0.4958
$(0\ 1\ 0)(0\ 1\ 1)_{12}$			0.3	0.0809	0.5916	0.5910	0.6912
			0.5	0.1483	0.6143	0.6111	0.7057
			0.8	0.3819	0.5848	0.5767	0.6732
$(0\ 1\ 1)(0\ 1\ 1)_{12}$		0.3	0.3	0.0561	0.5622	0.5629	0.6400
			0.5	0.1030	0.5788	0.5729	0.6541
			0.8	0.3201	0.5579	0.5527	0.6321
		0.5	0.3	0.0423	0.5554	0.5569	0.6080
			0.5	0.0844	0.5463	0.5341	0.6179
			0.8	0.2837	0.5333	0.5368	0.5917
		0.8	0.3	0.0255	0.5049	0.4971	0.5460
			0.5	0.0564	0.5102	0.5009	0.5484
			0.8	0.2189	0.4800	0.4789	0.5340
$(1\ 1\ 0)(0\ 1\ 1)_{12}$	0.3		0.3	0.1163	0.6033	0.6138	0.7678
			0.5	0.1942	0.6272	0.6320	0.7768
			0.8	0.4364	0.6163	0.6222	0.7449
	0.5		0.3	0.1436	0.6009	0.6016	0.7958
			0.5	0.2300	0.6284	0.6406	0.8046
			0.8	0.4759	0.6460	0.6510	0.7848
	0.8		0.3	0.1856	0.6208	0.6167	0.8336
			0.5	0.2882	0.6472	0.6432	0.8487
			0.8	0.5155	0.6664	0.6748	0.8261

Table 9: Average power for all spectrum peak diagnostics, using all 20 years.

used by Soukup and Findley (1999) to check for a trading-day effect was to see whether a "visually significant" peak was present in either the spectrum of the seasonally adjusted output series or the spectrum of the irregular component. This study considered each spectrum individually, and the spectrum of the model residuals appeared to perform the best as a stand-alone diagnostic. Of the four spectrum diagnostics, it had more consistent size values across models, whereas the irregular and seasonally adjusted spectra were more erratic. The diagnostic for the differenced original series was by far the worst of the four spectrum peak diagnostics. As was the case for Ladiray (2012), compared to the F-test, the spectrum peak diagnostics with the usual visual significance criterion have higher false detection rates, and with the criterion adjusted to produce a 0.05 false detection rate, they have considerably lower power.

By default, X-13ARIMA-SEATS uses only the last 96 observations to estimate the spectrum. For a quarterly series (not considered in this study), this amounts to 24 years worth of data, while it covers 8 years for a monthly series. This simulation study observes that the spectrum peak diagnostics for the model residuals are reasonably stable across the span used in estimating the spectrum, but that the power is noticeably higher when using a longer time span. For a fixed effect, this is not surprising. If the trading-day pattern for a series were to change over time in such a way that the shorter time span is not reflective of the rest of the series, this may not necessarily be the case.

A few alterations to the framework of the simulation study were not examined, but shall be briefly mentioned here. A normal distribution was enforced for the white noise process ξ_t in generating the simulated series. It might be useful to verify that the results observed in this

study are still valid in the event that a non-normal error distribution is encountered. Lytras et al. (2007) considered the effect of an incorrect model specification for detecting seasonality – a similar approach (or one in which X-13ARIMA-SEATS is allowed to choose the model) here might allow us to assess the impact on detecting a trading-day effect. Our simulation study adhered to monthly series of 20 years in length and spectrum spans of either 8 years or 20 years (full data). Although Ladiray (2012) did consider series of varying lengths, the observation was that the length of the series did not make a substantive difference in the overall results.

Acknowledgments The authors would like to thank William Bell and David Findley for helpful discussions in conducting this study.

References

Bell, W. R. and Hillmer, S. C. (1983), "Modeling time series with calendar variation," *Journal of the American Statistical Association*, 78, 526–534.

Cleveland, W. S. and Devlin, S. J. (1980), "Calendar effects in monthly time series: detection by spectrum analysis and graphical methods," *Journal of the American Statistical Association*, 75, 487–496.

Ladiray, D. (2012), "Theoretical and real trading-day frequencies," *Economic Time Series: Modeling and Seasonality*, 255–279.

Lytras, D. P., Feldpausch, R. M., and Bell, W. R. (2007), "Determining seasonality: A comparison of diagnostics from X-12-ARIMA," in *Proceedings of the International Conference on Establishment Surveys III.(CD-ROM)*.

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," in *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, pp. 144–149.

U.S. Census Bureau (2015), *X-13ARIMA-SEATS Reference Manual, Version 1.1*, U.S. Census Bureau, U.S. Department of Commerce, Washington, DC, http://www.census.gov/ts/x13as/docX13AS.pdf.