Journal of Official Statistics, Vol. 26, No. 2, 2010, pp. 371–394

Comparison of X-12-ARIMA Trading Day and Holiday Regressors with Country Specific Regressors

Christopher G. Roberts[1], *Scott H. Holan*[2], *and Brian Monsell*[3]

Several methods exist which can adjust for trading day and holiday effects in monthly economic time series. This article reviews and compares two such methodologies for conducting proper adjustments. The two methodologies are based upon the U.S. Census Bureau's X-12-ARIMA method and one developed by the Statistical Offices of the European Communities, commonly referred to as Eurostat. Three different methods are used to compare the U.S. Census Bureau procedure and the Eurostat-inspired procedure. These methods are spectral analysis, sample-size corrected AIC comparisons, and examination of out-of-sample forecast errors. Finally, these comparisons are conducted using nearly 100 U.S. Census Bureau time series of manufacturing data, retail sales, and housing starts along with roughly 70 Organisation for Economic Co-operation and Development (OECD) European time series of manufacturing, retail sales, and industry data. This empirical study is the first of its kind and therefore provides an important contribution to the seasonal adjustment community.

Key words: Eurostat; holiday effect; model selection; regARIMA model; trading day effect.

1. Introduction

Many time series are reported on a monthly basis and represent an aggregation of unobserved daily values. Since the daily values are unobserved, these particular time series often contain various elements that must be adjusted for in order to properly analyze the data. One of these elements is called a trading day effect (also called the day-of-week effect), which results from a combination of an underlying weekly periodicity in the unobserved daily data along with how many days of the week occur five times in a given month. For example, August of 2003 began on a Friday, so there were five Fridays, Saturdays, and Sundays in that month and only four of each of the other four days. In August of 2005, there were five Mondays, Tuesdays, and Wednesdays. Thus, the weekly periodicity combined with the differing numbers of each specific weekday will more than

[1] Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO, 65211-6100, U.S.A. Email: roberts.chrisg@gmail.com
[2] (to whom correspondence should be addressed) Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO, 65211-6100, U.S.A. Email: holans@missouri.edu
[3] Statistical Research Division, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100, U.S.A. Email: brian.c.monsell@census.gov
Acknowledgments: The authors would like to thank David Findley and William Bell for valuable discussion as well as Tom Evans for his assistance with various graphs and figures. Additionally, the authors would like to thank an associate editor and three anonymous referees for their comments which helped improve this article.
Disclaimer: This article is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

likely have a considerable effect on the time series, making it difficult to properly analyze the data unless these effects are adequately accounted for. A specific example of this occurrence is in ticket sales at a movie theater reported on a monthly basis. Sales are typically low at the beginning of the week and higher during the weekend. This weekly trend must be properly accounted for by trading day adjustment before a meaningful analysis of the data can be conducted. Additionally, holidays have a tendency to affect monthly time series. These elements, often called calendar effects, can be adjusted for using regARIMA models, which are regression models with seasonal ARIMA (autoregressive integrated moving average) errors. For a discussion of regARIMA models, see Bell (2004).

Methods used to adjust for trading day effects usually involve some form of counting the number of specific weekdays in a given month (i.e., the number of Mondays in January 2008, the number of Tuesdays in January 2008, . . . , the number of Sundays in January 2008) and then using these values as regressors. The U.S. Census Bureau has a particular procedure that it uses in its seasonal adjustment program, X-12-ARIMA (U.S. Census Bureau 2007). This procedure produces six distinct trading day regressors that can be expressed as

$$TD_{j,t} = D_{j,t} - D_{7,t} \tag{1}$$

where $D_{j,t}$ is the number of days in month t for weekday j, for $j = 1, \ldots, 7$ where 1 corresponds to Monday, 2 corresponds to Tuesday, . . . , 7 corresponds to Sunday.

A justification of these regressors is rather straightforward. Assuming that each day of the week has a fixed effect (or contribution), say α_j, we can write the overall effect of a particular month t as $\sum_{j=1}^{7} \alpha_j D_{j,t}$. This can be rewritten as the sum of two values, $\bar{\alpha}\sum_{j=1}^{7} D_{j,t}$ and $\sum_{j=1}^{7}(\alpha_j - \bar{\alpha})D_{j,t}$, for $\bar{\alpha} = 1/7\sum_{j=1}^{7}\alpha_j$. The value $\bar{\alpha}\sum_{j=1}^{7} D_{j,t}$ corresponds to a length of month effect. The length of month effect is handled in two ways. For non-February months the effect is automatically absorbed into the seasonal component of the decomposition of the series because these months have constant month-lengths. For February, the length of month effect is handled in preadjustments of the data or with a leap year regressor. Therefore, we can ignore the length of month effect and we are then left with $\sum_{j=1}^{7}(\alpha_j - \bar{\alpha})D_{j,t}$. This is the sum of the number of weekdays of a month times the particular weekday's deviation from the average daily effect $\bar{\alpha}$. Let $\beta_j = (\alpha_j - \bar{\alpha})$; noticing that $\sum_{j=1}^{7}\beta_j = 0$, it follows that $\beta_7 = -\sum_{j=1}^{6}\beta_j$ and hence

$$\sum_{j=1}^{7}(\alpha_j - \bar{\alpha})D_{j,t} = \sum_{j=1}^{7}\beta_j D_{j,t} = \sum_{j=1}^{6}\beta_j D_{j,t} + \beta_7 D_{7,t}$$

$$= \sum_{j=1}^{6}\beta_j D_{j,t} - \sum_{j=1}^{6}\beta_j D_{7,t} = \sum_{j=1}^{6}\beta_j (D_{j,t} - D_{7,t})$$

The values $(D_{j,t} - D_{7,t})$ for $j = 1, \ldots, 6$ are the six regressors defined in (1), with coefficients β_j corresponding to the deviation of the daily contribution of weekday j from the average daily effect $\bar{\alpha}$. Thus, the coefficients calculated in the X-12-ARIMA program are the β_js, where negative values correspond to weekdays with a smaller than average contribution and positive values correspond to weekdays with a larger than average contribution.

In addition to addressing trading day effects, X-12-ARIMA is capable of handling moving holiday effects through the inclusion of regressors for Easter Sunday, Labor Day, and Thanksgiving Day. These holidays are considered moving holidays because their effects on series have the potential to affect more than one month. The regressors each assume that the fundamental structure of the time series changes for a fixed number of days before each of these three holidays. Beginning on Easter Sunday and Labor Day, the nature of the time series returns to normal. For Thanksgiving Day, the fundamental structure of the time series remains altered until December 24th. Regressors for Labor Day and Thanksgiving Day are occasionally needed in the regARIMA model because of the effect these holidays sometimes have on retail sales data and other economic series for a number of days that extends into a second month. Additionally, a regressor for Easter is often necessary because the date of Easter Sunday occurs anywhere from March 22nd to April 25th in a given calendar year. The effects of other holidays, such as Martin Luther King, Jr. Day and Christmas Day, are believed to be absorbed by the seasonal component of the series because they are fixed (or stationary) on a particular date or a particular day of a given month and do not typically affect other months. Owing to their fixed nature, there is no need to include regressors for these holidays in a regARIMA model because they will be handled in the seasonal ARIMA component.

X-12-ARIMA also uses another, more parsimonious, model which was originally suggested by TRAMO (Gómez and Maravall 1996). This approach reduces the number of trading day regressors from six to one by assuming the daily effect of weekdays (Monday through Friday) is the same, and the daily effect of weekend days (Saturday and Sunday) is the same. Thus, the number of weekend days is subtracted from the number of weekdays, providing a single regressor; this regressor can be expressed as follows

$$TD1_t = \sum_{j=1}^{5} D_{j,t} - \frac{5}{2}\sum_{j=6}^{7} D_{j,t} \tag{2}$$

This regressor is derived in a similar fashion as the regressors from (1) with the constraints $\alpha_1 = \cdots = \alpha_5$ and $\alpha_6 = \alpha_7$. While this constrained model has fewer regressors, it is potentially less precise due to its fundamental assumption that weekdays have the same effect, and Saturdays and Sundays have the same effect.

Another approach to adjusting for trading day effects, that has been considered by Eurostat (the Statistical Office of the European Communities), does not make the assumption that fixed holidays are absorbed by the seasonal component. Instead, their method constructs a nominal count of days that accounts for fixed holidays. Although the Demetra 2.0 User Manual Release Version 2.0 (Statistical Office of the European Communities 2002) contains details regarding software implementation for the Eurostat method, no published study or description of the Eurostat method exists. In this direction, this article provides an explicit description of the method along with an extensive empirical study investigating its performance relative to the U.S. Census Bureau's X-12-ARIMA method.

Specifically, the regressors for the Eurostat methodology are composed by adding the number of holidays that fall on a specific day of the week (Monday, etc.) in a given month to the number of Sundays in the above-mentioned regressors for X-12-ARIMA.

We applied this European holiday count method to U.S. holidays (and country specific holidays). For example, in January of 2008 there are four Mondays. However, since Martin Luther King, Jr. Day falls on Monday, January 21st of 2008, the nominal number of Mondays is three. The Monday of Martin Luther King, Jr. Day is then added to the number of Sundays in the month. It is important to point out that Easter Sunday would not be taken into account by these regressors because it is already a Sunday. Furthermore, the effect that Easter has on the days preceding Easter Sunday is handled with its own moving holiday regressor. The fixed holidays that are accounted for in these modified trading day regressors are typically country specific and are used to adjust economic data produced in that particular country. Therefore, different countries with different holiday calendars will formulate regressors with different values. The motivation behind using such a method is that the various European countries have very different holiday calendars, making it difficult to compare economic data across countries when country specific holidays are not properly accounted for. These country specific trading day regressors, corrected for fixed holidays, can be expressed as

$$EU_{j,t} = (D_{j,t} - H_{j,t}) - (D_{7,t} + H_{j,t}) \tag{3}$$

where $H_{j,t}$ is the number of fixed holidays that fall on days $j = 1, \ldots, 6$. For these regressors, a holiday falling on a jth day of the week would be accounted for in the Sunday component of the regressor only for that particular jth day regressor. For the example of January 2008, we would have $EU_{1,t} = (4 - 1) - (4 + 1) = -2$, which is in essence 3 Mondays minus 5 Sundays. On the other hand, for Wednesdays we would have $EU_{3,t} = (5 - 0) - (4 + 0) = +1$, which is the same as $TD_{3,t} = 5 - 4 = +1$ for the Census Bureau regressors since there are no Wednesday holidays in January 2008.

Similar to the U.S. Census Bureau procedure, a simplified model can be described for the Eurostat procedure. This particular model modifies the regressor from (2) by considering holidays falling from Monday to Friday as Saturdays/Sundays. The regressor for this model can be expressed as follows

$$EU1_t = \sum_{j=1}^{5} (D_{j,t} - H_{j,t}) - \frac{5}{2}\left[\sum_{j=6}^{7} (D_{j,t}) + \sum_{j=1}^{5} (H_{j,t})\right] \tag{4}$$

A natural question that arises when considering the different methods that are used when seasonally adjusting monthly flow data is "Which method is better for a particular series, or even for a group of series?". In the context of this article the question becomes "Does the method currently employed by the U.S. Census Bureau do a better or worse job of seasonally adjusting monthly economic flow data than the Eurostat method?". Soukup and Findley (2000) describe three methods that can be used to compare the effectiveness of different models in properly accounting for trading day and holiday effects when seasonally adjusting monthly data, namely spectral analysis, comparison of modified AIC values, and analysis of out-of-sample forecast errors. These methods are employed here, using a collection of economic time series from the U.S. Census Bureau and the OECD, to compare the U.S. Census Bureau's X-12-ARIMA method of handling trading day and holiday effects with the Eurostat inspired method of using country specific regressors. After comparing these two methods, we determine which of the two is more effective in

adjusting for calendar effects in several separate groups of monthly economic flow series. Specifically, the series considered here are manufacturing series, industry series, retail series, and housing starts.

In Section 2 we detail the methods of analysis used on the collection of economic time series. In Section 3 we describe the specific nature of our analyses, including a description of our data, how we structured our models, and how the models were fit to our data. Section 4 contains a discussion on the implementation of our analyses and a summary of the results. Finally, Section 5 contains concluding remarks.

2. Methods of Analysis

As mentioned, the three methods of analysis used to determine which approach is better at adjusting for trading day and holiday effects are checking for visually significant trading day peaks in various spectra, comparing modified AIC values, and comparing out-of-sample forecast errors (OSFEs). The simplest method of analysis is done by examining three spectral plots for each model. The first spectrum is of the differenced, transformed, and seasonally adjusted series. The other two spectra are of the irregular series (also identified as the "final Henderson trend"-adjusted seasonally-adjusted series in X-12-ARIMA) and the residuals of the fitted series. Analysis of these spectra involves the identification of a significant spectral peak at a point along the spectrum that corresponds to trading day effects. For a spectrum $f(\lambda)$, $0 \leq \lambda \leq .5$, the two points that have been identified as those corresponding to trading day effects for monthly data are .348 and .432. The value .348 cycles/month comes from the number of weekly cycles that will occur in a month that has an average length. Due to leap year and the seven-day weekly cycle, the Gregorian calendar has a 28-year cycle. The average year is 365.25 days long, making the average month length $365.25/12 = 30.4375$ days. Thus, a week cycles through an average month $30.4375/7 = 4.348$ times, giving the fractional value of .348 when ignoring the ones unit to the left of the decimal point. Examining this particular peak has proven to be worthwhile in trading day adjustments (Soukup and Findley 1999). The other value, .432, was found to be important in detecting trading day effects by Cleveland and Devlin (1980). For a more comprehensive discussion regarding trading day frequencies, see Cleveland and Devlin (1980). Within X-12-ARIMA, a warning message is produced for any spectrum that has a "visually significant" peak at either of the two critical peaks associated with trading day effects (Findley, Monsell, Bell, Otto, and Chen 1998). The determination of a visually significant peak is done within X-12-ARIMA, but it is important to note that X-12-ARIMA currently has no method of testing a hypothesis for visual significance that is capable of producing a *p*-value associated with statistical significance (McElroy and Holan 2009).

An evaluation of two differing models using the spectral method of visual significance results in seeing which of the models being tested does not produce a visually significant peak on either of the spectra for a given series; or, perhaps, testing the models on a large number of series and seeing which model produces the least number of warning messages. Example spectral plots are provided in Figure 1. Specifically, Figure 1(a) displays the spectrum of a series that contains no visually significant trading day peaks after adjustments, whereas Figure 1(b) illustrates the spectrum of a series where trading day

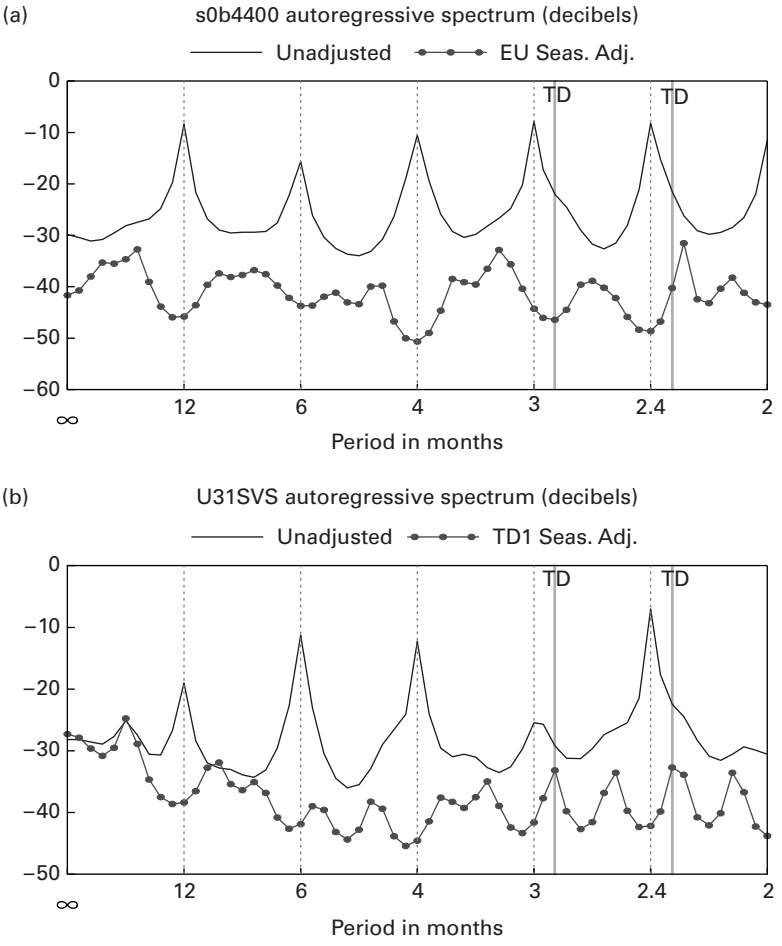

Fig. 1. Examples of spectral plots with identified trading day values

adjustment methods were unable to completely account for trading day effects, leaving visually significant peaks at important trading day values in the spectrum. For examples of spectral comparison plots (spectra of two models plotted together) see Figure 2. Obvious problems arise from this method of analysis, particularly that it does not allow for clear model selection when either both models produce warning messages or neither model produces warning messages. Additionally, because the spectra are assessed for visually significant peaks on an individual basis, this method of analysis contains no definable hypothesis test that can conclude a particular model is superior to another with any statistical significance.

To further facilitate the analysis of two separately proposed models for a particular series or a group of series, sample-size corrected Akaike Information Criterion (AICC) values are compared. AICC, proposed by Hurvich and Tsai (1989), is defined as

$$AICC_N = -2L_N + 2p\left(\frac{1}{1-(p-1/N)}\right) \tag{5}$$

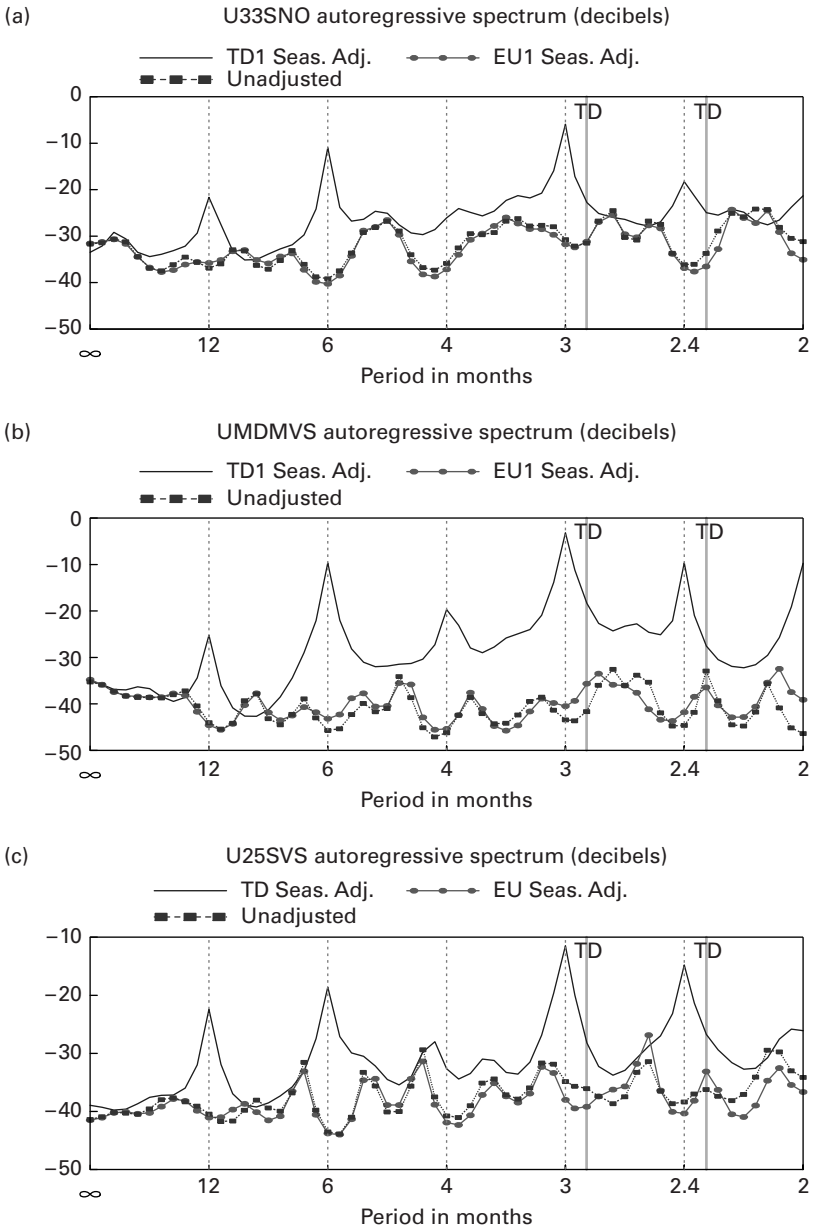

Fig. 2. Spectral comparison plots for trading day peaks, (a) has no peaks for either model, (b) has visually significant peaks for both models, (c) has peaks for both models, but only one is visually significant

where N is the number of observations, L_N is the maximized log-likelihood of an estimated regARIMA model fit to the N observations, and p is the number of parameters that are estimated in the model.

For some series, X-12-ARIMA detects additive outliers, level shifts, and temporary changes. After AICC values are computed, the model that does a better job of adjusting for calendar effects is determined by which one produces the lower AICC value. However,

two steps must be made in order for AICC values to be relevant when comparing two models applied to a series. The first is to make sure that both models use the same transformation and perform the same differencing operator on the series. The second is to have both models include the same outlier regressors. When an outlier is included in a particular model, it has a tendency to substantially increase the maximized log-likelihood, thereby lowering the observed AICC value. Thus, identified outliers can become a relatively large factor in model selection, as opposed to the more important data properties. In order to remedy this potential problem and to allow for an informative comparison of AICC values, it is necessary to make sure that both models include the same outlier regressors (X-12-ARIMA Reference Manual, U.S. Census Bureau 2007). Additionally, the magnitude of the difference between the two AICC values must be greater than 1.0 in order to consider one model superior to the other. Differences less than 1.0 in magnitude are considered to be inconclusive. AICC comparisons are commonly the primary method of analysis used in selecting a preferred model for individual series. For a comprehensive discussion on the appropriateness of this method see Burnham and Anderson (2002, 2004).

The third criterion used for model selection is to compare OSFEs. Out-of-sample forecasts are calculated for each model on a given series and have a particular lag associated with them. The two most common lags are 1 and 12, which respectively correspond to monthly differences and yearly differences. Calculation of out-of-sample forecasts is done using an iterative process. Specifically, a regARIMA model is fit to a sequence of data values contained within the series (perhaps the first 72 observations of a series that contains 96 total observations) and then a k-step-ahead forecast is made outside of the sequence. Subsequently, a regARIMA model is fit to a sequence of data values that are composed of the previous sequence with the addition of the next available data value. For example, monthly values from January 1995 to December 2003 are used to fit a regARIMA model and a forecast is made for January 2004. Then, the model is refit so that it accounts for values from January 1995 until January 2004, and a forecast is made for February 2004. This process would continue until forecasts were developed from January 2004 until the last month that had available data for the series. This particular example is of a 1-step-ahead forecast procedure; a similar procedure can be implemented for 12-step-ahead forecasts and would have given the first forecast for December 2004, the second for January 2005, and so on.

After this process is completed, OSFEs can be calculated by subtracting the observed values in the series from the forecasts. A special diagnostic can be created from these OSFEs that is able to compare exactly two separate models. This diagnostic uses the accumulated sum of squared OSFEs for each model, and then combines them in a normalized format to create a set of diagnostic values that can be plotted (e.g., Figure 3). A clearly visible upward or downward trend on this plot will reveal if the first or second model is more adequate. For a complete discussion, see Findley et al. (1998).

3. Data and Models

The data used for our analysis comes from two sources. The first source, provided by the U.S. Census Bureau, consist of monthly economic time series of manufacturing data, retail

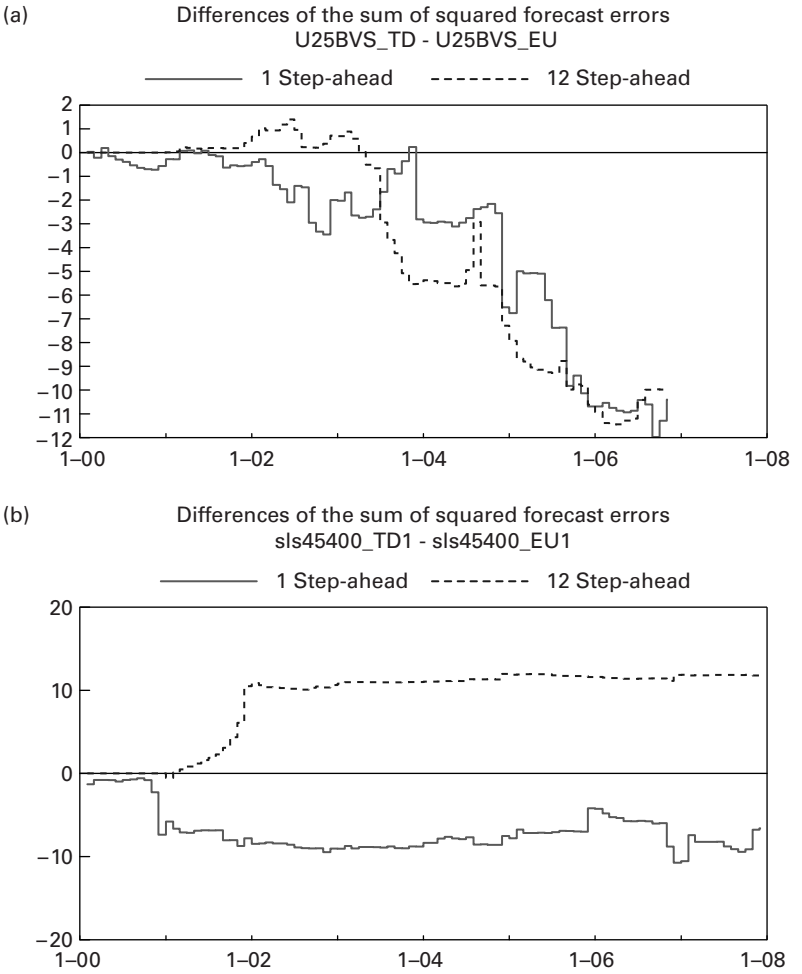

Fig. 3. Example OSFE comparison plots: (a) depicts a clear preference for the Census Bureau model on both lags, (b) preference is inconclusive

information, and housing starts. Information on the source and reliability of these series can be found at www.census.gov/cgi-bin/briefroom/BriefRm. The starting date for each series is either January or February of 1992; the ending date is November 2006 for all manufacturing series, December 2006 for all housing starts series, and December 2007 for all retail series. We conducted analyses on a total of 93 U.S. Census Bureau datasets; 54 were manufacturing series, 27 were retail series, and 12 were housing starts series.

The second set of series come from different European OECD member countries (Austria, Belgium, Denmark, Finland, France, Germany, Hungary, Italy, Netherlands, Norway, Spain, Sweden, and the United Kingdom) and consist of manufacturing, retail sales, and industry data. The starting date for each European series is January of 1992 and the ending date is either December 2007 or January 2008. We conducted analyses on a total of 69 European datasets; 35 were manufacturing series, 25 were retail series, and 9 were industry series.

For the remainder, the four models of concern will be referenced by the designation of their respective regressors in Formulas (1) through (4). Specifically they are models *TD*, *EU*, *TD*1, and *EU*1, where *TD* and *TD*1 are the U.S. Census Bureau models and *EU* and *EU*1 are the Eurostat-inspired models whose trading day regressors include a country specific fixed holiday correction.

For the U.S. Census Bureau time series, the regressors for the Eurostat-inspired models were created by accounting for the ten U.S. federally recognized holidays (excluding the quadrennial Inauguration Day). The ten federally recognized holidays are: New Year's Day, Martin Luther King, Jr. Day, Washington's Birthday, Memorial Day, Independence Day, Labor Day, Columbus Day, Veterans Day, Thanksgiving Day, and Christmas Day. It is important to note that, in Section 1, Labor Day and Thanksgiving were originally listed as moving holidays. However, we ran initial tests for their significance and found that the moving holiday regressors for those holidays were overwhelmingly showing up as not significant. Therefore, the only holiday treated as moving in our model was Easter. The federally recognized holidays (including Labor Day and Thanksgiving) used in the Eurostat models were strictly used as fixed holidays (that is, no additional regressors accounting for holiday effects in the days leading or trailing the holiday). In addition, this made the two models more comparable. The effect of the movement of Easter from year to year was handled with a separate regressor that was included in the models when it was found to be important through an AIC test within X-12-ARIMA. Just as Easter Sunday is not considered in the construction of the Eurostat regressors because it already is treated as a Sunday, any holidays with a fixed date (New Year's Day, Independence Day, Veterans Day, and Christmas Day) that happened to fall on a Sunday for a particular year were likewise not directly used in constructing the Eurostat regressors.

For our analysis, of the U.S. Census Bureau time series, we first compared the two Census Bureau models, *TD* and *TD*1, to determine whether a 6-regressor or 1-regressor model was more appropriate for each of the 93 series. This was done by first establishing whether or not a leap year adjustment was necessary for each series using the 6-regressor Census Bureau approach. The significance of leap year effects was determined within the X-12-ARIMA program using AlC-based selection criterion. The models *TD* and *TD*1 were then compared using AICC values to determine the more effective of the two. For this specific comparison, it is important to note that model *TD*1 is a nested case of *TD*, AICC differences are asymptotically equivalent to AIC differences and thus, for large enough series, vary approximately as a chi-square variate with degrees of freedom corresponding to the difference in the number of parameters for the two models *TD* and *TD*1. For a more detailed discussion, see the X-13A-S Reference Manual (U.S. Census Bureau 2008). After deciding whether six trading day regressors or a single trading day regressor should be used for each series, we compared the preferred Census Bureau model with its Eurostat-inspired counterpart. Thus, for each series either models *TD* and *EU* were compared, or models *TD*1 and *EU*1 were compared.

For our analysis, of the European time series, we first determined whether a 6-regressor or 1-regressor model would be more appropriate while concurrently determining if a leap year adjustment was necessary. In particular, this was done by comparing four separate models (with no outliers); these models were *EU* and *EU*1 with and without an adjustment for leap year. The model having the smallest AICC value was chosen. One exception to

this procedure occurred if the magnitude of the difference in the AICC values for *EU*-no leap year versus *EU*1-leap year or *EU*1-no leap year versus *EU*1-leap year was less than one. When this exception occurred, a leap year adjustment was included in the model for that particular series. It is important to note that this procedure differs from the procedure used for the U.S. Census Bureau data. Specifically, for the U.S. Census Bureau data, it was first determined whether a leap year adjustment was necessary using the 6-regressor Census Bureau approach. If a leap year adjustment was determined necessary we proceeded to a second stage of assessing whether a 6-regressor or 1-regressor *(TD* or *TD*1) model was more appropriate. This change in methodology, for the European series, reflects not assuming that the *TD* or *TD*1 models are the "default" models as was done for the U.S. series.

The datasets were analyzed in X-12-ARIMA, with a separate analysis being conducted for each of the two models chosen for every dataset (a Census Bureau model and its Eurostat counterpart for the primary analysis). A log transformation and differencing was carried out in each instance and leap year effects were taken into account when necessary. A multiplicative decomposition was assumed for all series. Specifically, we assumed

$$X_t = T_t * S_t * I_t * TD_t * H_t \tag{6}$$

for a series of observed values X_t with typical trend, seasonal, and irregular components T_t, S_t, and I_t, respectively. TD_t and H_t are the trading day and moving holiday components of the series. Note that for Eurostat methods TD_t would refer to the trading day regressors with the country specific adjustments of the fixed holidays. Estimates for all components of the series were made through the log-transformed and differenced series. The trading day and moving holiday elements were handled using regression techniques and errors of this regression fit were considered to be seasonal errors, which were modeled with a seasonal ARIMA (SARIMA) component. Again, the estimation is conducted within the X-12-ARIMA program using the *automdl* specification to automatically determine the proper $(p,d,q) \times (P,D,Q)_{12}$ SARIMA model for the monthly time series. Specifically, this amounts to using iterative generalized least squares (IGLS), with iteration occurring between the regression and ARMA parameters, (Otto et al. 1987; Bell 2004). Further, maximum order limitations of 3 and 2 were used on the regular ARMA and seasonal ARMA polynomials, respectively. For a complete discussion regarding SARIMA models, see Shumway and Stoffer (2006). Additionally, three different types of outliers were automatically detected: additive outliers, level shifts, and temporary change outliers. The overall goal is to properly adjust series for seasonality as well as trading day and holiday effects by estimating their respective components and then dividing them out of the decomposition so that the adjusted series only contains trend and irregular (error) components.

For all U.S. series, an Easter holiday regressor was considered for each model and would correspond to H_t in (6). According to the X-12-ARIMA Reference Manual (U.S. Census Bureau 2007), this regressor assumes that w days before Easter the "level of activity changes . . . and remains at the new level until the day before [Easter Sunday]." The general form of this regressor is called *easter*[w] in X-12-ARIMA, where w references the number of days before Easter that the shift occurs. The most commonly used values for w are 1, 8, and 15. For our analysis, *easter*[8] was used because of its

strength in accounting for more than simply the Saturday before Easter Sunday and because it has been shown to be more often preferred in previous studies (Findley and Soukup 2000). An AIC test was conducted within X-12-ARIMA to determine whether or not the *easter*[8] holiday regressor was necessary in the models. For more information on how Easter effects are handled in the X-12-ARIMA program, see the X-12-ARIMA Reference Manual (2007). If for a specific series one particular model required the *easter*[8] regressor but the other model did not, the regressor was included in both models. The series was then rerun through X-12-ARIMA for each method. This was done for reasons similar to combining outlier sets, specifically that the interest of our investigation is in how well two different types of regressors can handle trading day adjustments, and not how effective the two procedures are in identifying outliers or Easter effects. It should be noted that *easter*[8] might not be the most appropriate Easter regressor for every series used in our study. However, the use of a single value for w was done primarily to simplify generalized comparisons. Finally, it may be the case that the concept of Easter is inappropriate for some of the series being investigated. Therefore, as previously discussed, we rely on an AIC test to determine whether inclusion of an Easter regressor is warranted.

For the European time series, the regressors for the Eurostat-inspired models were created using the country specific holiday counts from the software Demetra 2.2 (available at http://.circa.europa.eu/Public/irc/dsis/eurosam/library?l=/software/demetra_software&vm=detailed&sb=Title), which contains default holiday calendars. This method of adjustment is consistent with ESS guidelines on seasonal adjustment (Eurostat Methodologies and Working papers 2009). The effect of Easter was handled in a same manner as the U.S. series. However, for countries observing Maundy Tuesday or Good Friday as default holidays only an *easter*[1] regressor was tested for inclusion. For all other countries an *easter*[8] regressor was tested analogous to the U.S. series. The appropriate *easter*[w] regressor was included in both the Eurostat and Census models (i.e., both *TD* and *EU* or both *TD*1 and *EU*1) and an AIC test for inclusion was conducted. If either of the Eurostat or Census models included *easter*[w], then it was included in both models for the remainder of the process.

A number of diagnostics were gathered in order to compare the U.S. Census Bureau and Eurostat-inspired procedures. The first of these was the visually significant peaks in various spectra, as described above. In order to create OSFE plots comparing the two methodologies, the evolving (or accumulated) sums of squared out-of-sample forecast errors were computed for 1-step-ahead and 12-step-ahead forecasts. For each of the two k-step-ahead procedures, the forecasting capabilities of the two models being compared was addressed by creating a standardized difference of the errors at each time point. In order to properly use AICC values for comparisons, outlier adjustments had to be made. If the outliers identified for the two models being compared were not all the same, then the dataset was run through X-12-ARIMA again for each of the two models with the complete list of combined outliers included with outlier regressors. The original SARIMA $(p,d,q) \times (P,D,Q)_{12}$ components from the initial X-12-ARIMA runs for each model were then used for the second run (the *automdl* specification was not used). This was done to ensure that AICC values could be properly compared and so the newly introduced outlier regressors would not be able to influence an *automdl* procedure from choosing a different SARIMA component.

To develop the diagnostics for OSFE comparisons of k-step-ahead forecasts of a particular time series X_t, for $k \geq 1$, we take interest in a regARIMA model of the transformed series $x_t = f(X_t)$. For the N data points of the series, we let N_0 be an integer less than $N - k$ that is large enough for the data x_t to assume reasonable estimates of the model's coefficients for $1 \leq t \leq N_0$. Establishing N_0 in such a way will ensure that the forecasts are derived from a reasonable model. Then, for each t in $N_0 \leq t \leq N - k$, let $x_{t+k|t}$ denote the forecast of x_{t+k} conditioned on the estimated regARIMA model using the data $x_{t'}$, $1 \leq t' \leq t$. The out-of-sample k-step-ahead forecast of X_{t+k} will be $x_{t+k|t} = f^{-1}(x_{t+k|t})$. The out-of-sample forecast error for time $t+k$ is defined as $e_{t+k|t} = X_{t+k} - X_{t+k|t}$. The accumulated (or evolving) sums of squared out-of-sample forecast errors, as reported in X-12-ARIMA, are

$$SS_{k,M} = \sum_{t=N_0}^{M} e_{t+k|t}^2, \quad M = N_0, \ldots, N-k$$

In order to compare two separate models with forecast errors $e_{t+k|t}^{(1)}$ and $e_{t+k|t}^{(2)}$ with sums of squared errors $SS_{k,M}^{(1)}$ and $SS_{k,M}^{(2)}$, we compute a normalized (standardized) diagnostic of the differences of $SS_{k,M}^{(1)}$ and $SS_{k,M}^{(2)}$. This diagnostic is defined as

$$SS_{k,M}^{1,2} = \frac{SS_{k,M}^{(1)} - SS_{k,M}^{(2)}}{SS_{k,N-k}^{(2)}/(N-k-N_0)} \tag{7}$$

for $N_0 \leq M \leq N - k$. The recursion formula for (7),

$$SS_{k,M+1}^{1,2} = SS_{k,M}^{1,2} + \frac{\left(e_{k+M+1|M+1}^{(1)}\right)^2 - \left(e_{k+M+1|M+1}^{(2)}\right)^2}{SS_{k,N-k}^{(2)}/(N-k-N_0)}$$

shows that a plot of this diagnostic, as a function of M, will reveal a possible preference of either Model 1 or Model 2 in terms of their forecasting abilities, depending on the direction of the plotted diagnostic. Examples of OSFE plots are shown in Figure 3.

When attempting to determine which method was generally preferred by OSFE diagnostics, there was an issue concerning the possibility that for a particular series a plot of one lag may favor one model while a plot of the other lag may favor the other model (or no model at all). When this situation occurred, a particular model was considered to be superior to another if the lag 12 plot favored the model and the lag 1 plot either favored the model or was unable to favor either of the two models. Whenever the directional trend of the two lags differed, the results were considered to be inconclusive. The superiority of lag 12 over lag 1 in determining forecasting capabilities was chosen based on considerations for X-11 seasonal adjustment.

Finally, these three diagnostics – visually significant peak counts, AICC values, and OSFE plots – were then analyzed and used to determine which methodology was preferred in adjusting for trading day and holiday effects for the various collections of series – manufacturing, retail, industry and housing starts. The information gathered was also used to determine a preferred model for individual series (results available on request). When comparing the two models, a great deal of weight was placed upon

the AICC comparisons and the OSFE diagnostics received slightly less weight in the decision-making process. Since the process of comparing visually significant peaks is not as rigorous as the other diagnostics, its role was less emphasized.

4. Implementation and Results

For the analysis of the U.S. manufacturing series 54 datasets were used. The datasets that were run through the X-12-ARIMA program revealed that neither methodology was completely preferential across all of the datasets, though an argument could be made in favor of the U.S. Census Bureau models. Of the 54 series, 13 were analyzed using the 6-regressor models and 41 were analyzed using the constrained single regressor models. As displayed in Table 1, a total of 13 series for the Census models and 10 series for the Eurostat models had visually significant trading day peaks in either of the three spectra analyzed. More specifically, models *TD* and *TD*1 left 4 and 9 series with trading day peaks and models *EU* and *EU*1 left 3 and 7 series with trading day peaks. If anything, the Eurostat methodology reveals a very slight advantage over the U.S. Census Bureau models, but the difference between the two is small enough that no definitive preference can be established for the entire group of series on the basis of this analysis alone.

Whereas the visually significant peak counts revealed no preference, the AICC comparisons for the U.S. manufacturing series – made after appropriate outlier adjustments – point towards the Census method being the more appropriate way to adjust for trading day and holiday effects. A total of 28 series favored the Census models, whereas 15 series favored the Eurostat models with country specific holiday correction regressors (see Table 2). Furthermore 11 of the 54 series yielded AICC values that were too close to determine a preferred model, that is, their magnitude of difference was less than 1.0. Additionally, only 5 of the 43 conclusive series had AICC differences greater than 10.0 in magnitude. Nevertheless, it appears that AICC comparisons reveal a moderate preference for the U.S. Census Bureau procedure over the Eurostat procedure for the U.S. manufacturing series considered in our analysis. Although it may be possible to conclude "overall" preference for one specific method, in practice one would want to use whichever method is best for a particular series.

Table 1. Numbers of series with visually significant trading day peaks in the plots of the default (IRR and SA) and residuals spectra before outlier adjustments

	Trading Day Peaks – U.S. Census Bureau Series					
	Series and models					
	Manufacturing		Retail		Housing starts	
Source of warnings	Census	Eurostat	Census	Eurostat	Census	Eurostat
IRR or SA, and Residual Spectra	2	1	2	4	0	1
IRR or SA Spectra only	11	6	6	2	2	2
Residual Spectrum only	0	3	4	4	0	2
Totals	13	10	12	10	2	5

Table 2. Number of series favored when comparing TD with EU, and when comparing TD1 with EU1. SARIMA components were both hardcoded before outlier sets were joined together and decided by automdl *after outlier sets were joined together*

	AICC Preferences – U.S. Census Bureau Series					
	"hardcoded" before outlier			"*automdl*" after outlier		
	Census	Eurostat	Indifferent	Census	Eurostat	Indifferent
Manufacturing Series	28	15	11	9	18	27
Retail Series	20	3	4	16	5	6
Housing Starts Series	11	1	0	4	3	5

With the more subjective OSFE comparisons, the analysis revealed no clear preference for either of the two methodologies for the U.S. manufacturing series. As outlined in Table 3, 2 series favored the U.S. Census Bureau models and 4 series favored the Eurostat models, leaving 48 series with no conclusive results. Though the Eurostat method was slightly favored by OSFE comparisons over the Census method, we find that an overwhelming number of series produced OSFE plots that had no clear directional trend. Nevertheless, it is worth noting that when referring only to lag 1 performance, Census models were favored by 11 series and Eurostat models were favored by only 3 series.

Concerning the set of monthly economic series of U.S. manufacturing data, it is not possible from our analysis to definitively establish which of the two methodologies, the Census Bureau or the Eurostat, is more effective in adjusting for trading day and holiday effects across the entire group of series. However, a comparison of AICC values reveals a preference for Census Bureau procedures, while plots of sums of squared out-of-sample forecast errors depict no general preference. In addition, counts of visually significant trading day peaks in seasonally adjusted, modified irregular, and residuals spectra point towards inconclusiveness. This inconclusiveness further illustrates the importance of using whichever method is best for a particular series in practice.

Since 15 of the 54 U.S. manufacturing series preferred a Eurostat model with regard to AICC values, it seems that it would be beneficial to individually examine these series to see whether or not the other methods of analysis supported this conclusion. Taking the

Table 3. Number of Manufacturing, Retail and Housing Starts series favored by OSFE plots

	OSFE Model Preference – U.S. Census Bureau Series					
	Series and models					
	Manufacturing		Retail		Housing starts	
	Census	Eurostat	Census	Eurostat	Census	Eurostat
Lag 1 Preference	11	3	11	1	3	0
Lag 12 Preference	2	5	7	6	1	0
General Preference	2	4	7	3	1	0

AICC results to be the primary deciding factor in selecting a preferred model, we examined the OSFE plots and spectral plots on an individual basis to determine if these methods of analysis presented any reason to refute the AICC results. In only 1 of the 15 series was there any evidence to contradict the AICC comparisons. In this particular case, a lag 1 OSFE plot clearly depicted a preference for the U.S. Census Bureau's model. The three spectra contributed no conclusive evidence one way or the other. Thus, considering AICC values to be the primary benchmark in selecting model preferences for individual series, there is evidence that 14 of the manufacturing series preferred the Eurostat methodology over the U.S. Census Bureau methodology in adjusting for trading day and holiday effects. A similar examination of individual series where AICC comparisons favored models *TD* and *TD*1 revealed that 26 series preferred the Census models. Thus, for the U.S. manufacturing series, being examined on an individual basis, 26 preferred Census models, 14 preferred Eurostat models, and 14 comparisons proved inconclusive.

The U.S. retail series were entirely preferential towards the U.S. Census method. Of the 27 series examined, 25 were compared using models *TD* and *EU*, and 2 were compared using models *TD*1 and *EU*1. The overwhelming preference for the 6-regressor models is to be expected due to the large influence of weekly cycles in retail sales. The counts of visually significant trading day peaks reveal that both methods produced roughly the same number of warning messages. The U.S. Census Bureau models produced 12 and the Eurostat models produced 10. A closer inspection of the AICC values significantly helps strengthen the case for superiority of models *TD* and *TD*1 over *EU* and *EU*1. Table 2 reveals an overwhelming preference for the current X-12-ARIMA method of handling trading day effects in retail series. The comparisons favored U.S. Census Bureau models for 20 of the 27 series, only 3 series favored the Eurostat models, and 4 series produced AICC values with differences less than 1.0 in magnitude. This evidence definitively indicated the superiority of the U.S. Census Bureau's methodology over the Eurostat methodology when it comes to the U.S. retail series examined here.

Observing the sums of squared residual plots produced from out-of-sample forecasting provides additional evidence for models *TD* and *TD*1. As shown in Table 3, only 3 of the 27 comparisons showed a distinct preference for the Eurostat methodology, while 7 comparisons revealed a preference for the U.S. Census Bureau models, particularly the 6-regressor model. Even though the OSFE plots produced 17 of 27 inconclusives, the U.S. Census Bureau's current X-12-ARIMA models were superior to their Eurostat-inspired counterparts in terms of out-of-sample forecasting by a margin of more than 2 to 1. Additionally, when looking at only lag 1 OSFE plots it is clear that the U.S. Census Bureau models are preferred, with 11 series favoring models *TD* and *TD*1 and only 1 series favoring the Eurostat models.

Overall, the three methods of counting visually significant trading day peaks, comparing AICC values, and examining OSFE diagnostics revealed a strong preference for models *TD* and *TD*1 over their respective Eurostat counterparts when applied to U.S. retail sales data. Though visually significant peak counts did not favor one methodology over the other, examining the AICC comparisons indicated that the majority of the U.S. retail series produce smaller AICC values for the U.S. Census Bureau procedures than those produced under the Eurostat models. In addition, even though the OSFE plots were not as definitive as the AICC comparisons, they still indicate superiority of the U.S. Census Bureau's

methodology. Upon examining the 27 U.S. retail series on an individual basis, we found that there was not a single series where OSFE plots and spectral plots called into question the model selected by way of AICC comparisons. Thus, 3 of the U.S. retail series were preferential towards the Eurostat models whereas 20 series favored a U.S. Census Bureau model, with 4 series yielding inconclusive results.

Analysis of the 12 U.S. housing starts series revealed support for using the current U.S. Census Bureau models. For these series, 3 were examined using models *TD* and *EU*, while 9 were examined with 1-regressor models. Visually significant trading day peaks were found in 5 of the 12 series when models *EU* and *EU*1 were applied; the U.S. Census Bureau models produced a total of 2 series with visually significant trading day peaks.

AICC comparisons yielded a similar preference for the methodology currently used by the U.S. Census Bureau Models *TD* and *TD*1 were favored in 11 of the 12 series, with the remaining series favoring a Eurostat-inspired model. OSFE plots were entirely inconclusive (see Table 3). Only 1 series revealed a distinct preference for a particular model, and that was for a Census model. Both lag 1 and lag 12 OSFE plots for this series were conclusive. The lag 1 plots of 2 other series revealed a preference for the U.S. Census Bureau method, but lag 12 plots were inconclusive in each case.

Overall, analysis of the U.S. housing starts series indicates a preference for the U.S. Census Bureau methodology. However, the overwhelming indifference when examining forecasting capabilities is something that should not be overlooked. If a model's ability to accurately predict future occurrences is considered to be an important quality, then it would be hard to firmly declare that models *TD* and *TD*1 were more effective in handling trading day and holiday effects than their Eurostat counterparts. On the other hand, if a lower AICC value is the primary focus in model selection, then there is sufficient evidence supporting the use of current U.S. Census Bureau models on housing starts series. Using AICC results as the criterion for model selection on the 12 series individually, and examining OSFE and spectral plots as a means to support or refute AICC results, we end up with 11 series preferring the Census models and 1 series preferring the Eurostat models. The only series preferring country specific regressors exhibited an AICC reduction of only 1.49 in comparison to the competing U.S. Census Bureau model.

There is one very important observation to note regarding AICC comparisons for all of the U.S. series. The results of the comparisons presented here were done in such a way that the SARIMA components of the models determined with the *automdl* specification in X-12-ARIMA were used, then combined outlier sets were included and the SARIMA model was not allowed to change. This was done in order to prevent additional outliers from potentially distorting the seasonal ARIMA components. When SARIMA components were left up to *automdl* during the inclusion of combined outlier sets, there was a substantial alteration in the AICC comparisons, as shown in Table 2. Particularly, a sizable number of series produced AICC values that were very close for both Census and Eurostat models. Overall, the number of inconclusive series (AICC differences being less than 1.0 in magnitude) went from 15 to 38 for the 93 series examined. Moreover, AICC comparisons for the U.S. housing starts series were no longer able to conclusively indicate preference toward the Census methodology as being more appropriate, and the number of inconclusive comparisons for the U.S. manufacturing series increased to 27 of 54 total series. Alterations in visually significant peak counts and OSFE plots were minimal.

The sensitivity of the AICC values to the additional inclusion of outliers and the revision of the SARIMA component (through the *automdl* specification) seems to indicate that the two methodologies, overall, have a very similar effect on trading day adjustments of economic flow series. Clearly there are some series that definitively prefer one method over the other, but there are also a significant number of series for which comparisons can be considered inconclusive.

For the analysis of the European manufacturing series 35 datasets were used. Of the 35 series, 13 were analyzed using a 6-regressor model whereas 22 were analyzed using a constrained single regressor model. An examination of the spectra of the 35 datasets revealed a modest preference for the U.S. Census procedures. As displayed in Table 4, a total of 11 series for the U.S. Census models had visually significant trading peaks in at least one of the three spectra analyzed, whereas 17 series had visually significant trading peaks for the Eurostat models. More specifically, models *TD* and *TD*1 left 4 and 7 series with trading day peaks and models *EU* and *EU*1 left 7 and 10 with trading day peaks. Thus, on the basis of trading day peaks, the U.S. Census Bureau models displayed a slight advantage.

In contrast to the visually significant trading day peaks, results of the AICC comparisons revealed a clear preference for the Eurostat models (Table 5). In particular, of the 35 European manufacturing series, 26 had lower AICC values with the Eurostat models whereas only 8 series had lower AICC values with the U.S. Census models and 1 series had an AICC difference that was less than 1.0 in magnitude.

Examination of the OFSE plots for the European manufacturing series revealed no clear preference for either of the two methodologies. As depicted in Table 6, only 10 of the 35 series exhibited a distinct preference for either model, 5 preferring a U.S. Census model and 5 preferring a Eurostat model. In addition, when only evaluating lag 1 OSFE performance Eurostat models are preferred, with 15 series favoring *EU* and *EU*1 and only 3 series favoring the U.S. Census models.

Using AICC differences as a primary deciding factor of model preference, and using OSFE plots and spectral plots as a means to support or refute the AICC comparisons, 4 of the 26 series preferring a Eurostat model displayed evidence to contradict the AICC preferences. In each case, the lag 12 OSFE plot clearly depicted a preference toward the

Table 4. Numbers of series with visually significant trading day peaks in the plots of the default (IRR and SA) and residuals spectra before outlier adjustments

	Trading Day Peaks – European Series					
	Series and models					
	Manufacturing		Retail		Industry	
Source of warnings	Census	Eurostat	Census	Eurostat	Census	Eurostat
IRR or SA, and Residual Spectra	3	3	1	2	0	3
IRR or SA Spectra only	5	9	3	6	4	1
Residual Spectrum only	3	5	5	3	3	1
Totals	11	17	9	11	7	5

Table 5. Number of series favored when comparing TD with EU, and when comparing TD1 with EU1. SARIMA components were both hardcoded before outlier sets were joined together and decided by automdl *after outlier sets were joined together*

	AICC Preferences – European Series					
	"hardcoded" before outlier			"*automdl*" after outlier		
	Census	Eurostat	Indifferent	Census	Eurostat	Indifferent
Manufacturing Series	8	26	1	8	26	1
Retail Series	12	11	2	12	12	1
Industry Series	2	5	2	2	5	2

U.S. Census models. A similar analysis of the 8 series where the U.S. Census models produced lower AICC values resulted in only 1 series having evidence to refute the AICC preference. Specifically, the lag 12 OSFE plot supports the contrasting Eurostat model. Thus, on an individual basis, 22 series preferred a Eurostat model, 7 preferred a Census model, and 6 comparisons proved inconclusive.

For the analysis of the European retail series, 25 datasets were used. Out of the 25 series, 19 were analyzed using the 6-regressor models and 6 were analyzed using the constrained single regressor models. Both the U.S. Census Bureau and the Eurostat-inspired methods produced roughly the same number of warning messages for visually significant trading day peaks (Table 4). The U.S. Census models produced 9 warnings messages and the Eurostat models produced 11.

Analysis of AICC differences revealed similar inconclusiveness. U.S. Census models were preferred in 11 series and Eurostat models were preferred in 10 (Table 5). Furthermore, the OSFE plots revealed a distinct preference for the U.S. Census approach. In particular, 5 series showed a clear preference for the Census models and only 1 series preferred the Eurostat models, with the remaining 18 series being inconclusive.

On an individual basis, 5 of the series had OSFE or spectral plots that contradicted an AICC preference. Specifically, 3 of these series corresponded to Eurostat models with a smaller AICC value. Of these series, 2 had a lag 12 OSFE plot preferring the contrasting U.S. Census model while the third series had a lag 1 OSFE plot preferring the contrasting U.S. Census model. The remaining 2 series had AICC differences preferring a U.S. Census

Table 6. Number of Manufacturing, Retail and Industry series favored by OSFE plots

	OSFE Model Preference – European Series					
	Series and models					
	Manufacturing		Retail		Industry	
	Census	Eurostat	Census	Eurostat	Census	Eurostat
Lag 1 Preference	3	15	6	4	2	3
Lag 12 Preference	7	5	5	4	3	0
General Preference	5	5	5	1	3	0

model, but the lag 12 OSFE plots refuted that choice. Thus, on an individual basis, 10 series preferred a U.S. Census model, 7 preferred a Eurostat model and 7 were inconclusive.

There were only 9 European industry series analyzed. Two were analyzed using the 6-regressor models and 7 were analyzed using a constrained single regressor model. Five series had visually significant trading day peaks in the spectra for the Eurostat models whereas the U.S. Census models had 7 (Table 4).

AICC comparisons revealed 5 series that preferred a Eurostat-inspired model, 2 series that preferred a U.S. Census model and 2 series that had AICC differences less than 1.0 in magnitude (Table 5). There were only 2 series that favored the U.S. Census methodology: the AICC differences were 78 and 124.2 respectively, giving an average of 101.1. This is larger than expected since most AICC differences were less than or equal to 20 in magnitude. OSFE plots revealed a U.S. Census model preference for 3 series while the Eurostat model was not preferred in any series.

On an individual basis, there was 1 series with AICC differences revealing preference for a Eurostat model where a lag 12 OSFE plot showed a distinct preference for the U.S. Census model. All other series had no evidence in the OSFE or spectral plots that refuted the AICC preference. Thus, the Census method was preferred in 2 series, the Eurostat method was preferred in 4 series, and the remaining 3 series were inconclusive.

Of the 69 European series, 27 had visually significant trading day peaks in either of the residual, seasonally adjusted, or modified irregular spectra for the U.S. Census models and 33 series had visually significant trading day peaks for the Eurostat-inspired models. AICC differences revealed a general preference for Eurostat models over their Census counterparts. Specifically, 44 series preferred Eurostat models, 22 series preferred U.S. Census models and 3 additional series had AICC differences less than 1.0 in magnitude. Of the 13 countries examined, only the United Kingdom had more series prefer U.S. Census models to Eurostat models, 5 series favored the U.S. Census methodology and no series favored the Eurostat methodology.

For the analysis of the European series, to compare AICC values, the combined outlier sets were included in the models and the SARIMA model was not allowed to change. When SARIMA components were left up to *automdl* during the inclusion of combined outlier sets, there was no qualitative difference in the results for the 69 European series. This is a stark contrast to the results of the U.S. series, where *automdl* significantly increased the number of series with inconclusive model comparisons.

When examining the preferences of the series, an interesting question arises concerning the clear preference of the U.S. Census Bureau models by the U.S. retail series. Considering that U.S. manufacturing and housing starts series did not indicate as strong of a preference for either methodology and that both sets of series overwhelmingly favored the single regressor models, was the definitive preference seen in the retail series a result of the frequent use of the 6-regressor models? Considering the way in which the regressors for the Eurostat models augment the regressors currently used in X-12-ARIMA, it is plausible that a misspecification in the nature of fixed holidays would be more easily detected in models containing six regressors as opposed to those with a constrained single regressor. If all holidays do not act as Sundays – as is assumed in the Eurostat-inspired models with country specific trading day regressors – then this misclassification could

potentially affect all 6 regressors and their estimated coefficients. In order to rectify this potentially hazardous problem, it would be necessary to individually consider the nature of every holiday and determine whether it should be classified as a Sunday, or perhaps as a Friday instead. The uniqueness of individual series would require a subjective assessment of this issue for each series on an individual basis. An accurate reallocation of holidays could possibly improve the model's ability to adjust for trading day and holiday effects, making it more effective than the current U.S. Census Bureau procedure for a wide range of series. However, a technique of this sort is obviously prohibitively time consuming. Nevertheless, evidence has shown that a number of series examined here exhibit improvements over the current U.S. Census Bureau procedures in adjusting for trading day effects when the use of country specific fixed holiday corrections are employed. It is entirely plausible that a revised treatment of fixed holidays for some individual series could improve AICC values and possibly even OSFE diagnostics by enough of a margin so as to consider the Eurostat-inspired models capable of outperforming current U.S. Census Bureau models for the U.S. series. As the country specific regressors currently stand, however, only 17 of the 93 U.S. series have been shown to prefer this particular method over the current U.S. Census Bureau method of assuming the effect of fixed holidays to be entirely contained within the seasonal component of the series.

5. Concluding Remarks

We have reviewed and provided a detailed comparison of two differing approaches to adjusting for trading day and holiday effects in monthly economic time series. The first approach was the current Census Bureau procedure of creating trading day regressors in X-12-ARIMA, which assumes that the effects of fixed (stationary) holidays are absorbed by a seasonal component. The second approach was inspired by Eurostat, wherein fixed holiday effects are not assumed to be absorbed by a seasonal component and are instead accounted for by including fixed holiday counts in the calculation of X-12-ARIMA trading day regressors and treating holidays as Sundays. The methods used to compare these two approaches were a spectral analysis where visually significant trading day peaks were documented and compared, a comparison of AICC values after outlier adjustments, and an analysis of out-of-sample forecasting capabilities through plots of a standardized diagnostic of the differences of sums of squared out-of-sample forecast errors.

We have seen that counts of visually significant trading day peaks in the seasonally adjusted, modified irregular, and residuals spectra were roughly the same for each of the Census Bureau and Eurostat-inspired methods for each of the three groups of U.S. series analyzed. For the U.S. series, AICC comparisons revealed a preference towards the Census approach for all three groups of series. Box plots of AICC values by models used are displayed in Figure 4. However, an alternate examination of AICC values significantly increased the number of inconclusive comparisons for all three groups of U.S. series, particularly the U.S. manufacturing series. Further, for the U.S. series, OSFE plots tended to favor the U.S. Census Bureau models over their Eurostat-inspired counterparts, though an overwhelming number of plots did not favor either approach. There is evidence that specific groups of series may definitively prefer one approach over the other, such as the clear preference towards the current X-12-ARIMA approach for the U.S. retail series.

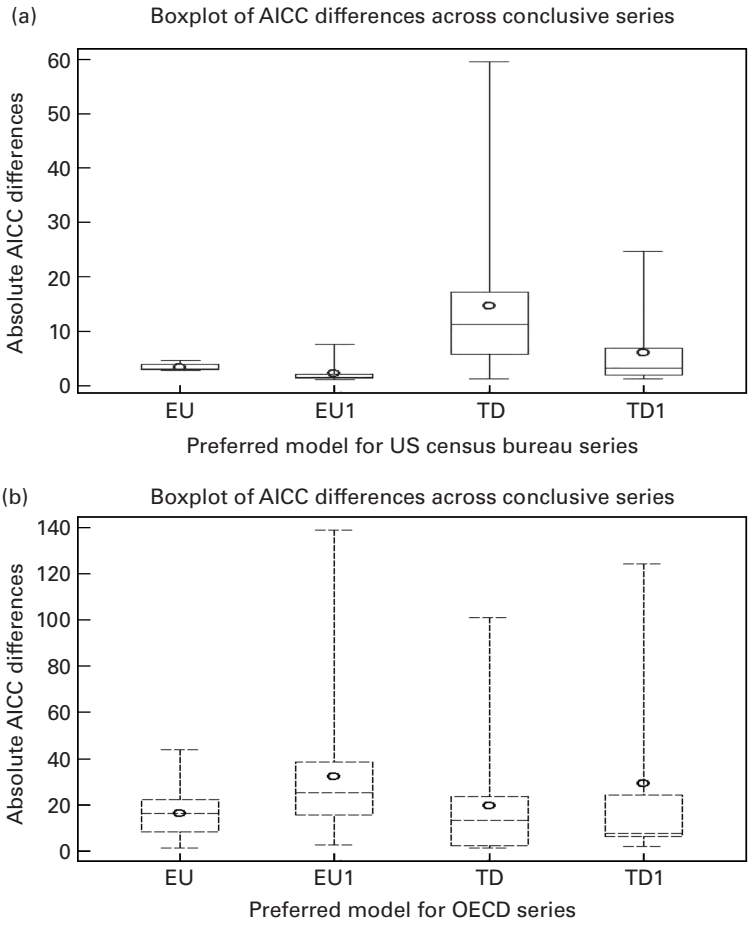

Fig. 4. (a) Box plots of absolute AICC difference values by model for 78 U.S. series with AICC differences greater than 1.0 in magnitude. The number of series that preferred EU was 6, EU1 was 13, TD was 30, and TD1 was 29. (b) Box plots of absolute AICC difference values by model for 52 European series with AICC differences greater than 1.0 in magnitude. The number of series that preferred EU was 15, EU1 was 18, TD was 11, and TD1 was 8

Similarly, we have seen that counts of visually significant trading day peaks in the seasonally adjusted, modified irregular, and residuals spectra were roughly the same for each of the Census Bureau and Eurostat-inspired methods for each of the three groups of European series analyzed. For the European series, AICC comparisons revealed a preference toward the Eurostat approach for the manufacturing series. However, for the retail and industry series no clear preference can be gleaned. Box plots of AICC values by models used are displayed in Figure 4. Additionally, alternate examination of AICC produced similar results for all three groups of European series. Further, for the European series, OSFE plots tended to favor the U.S. Census Bureau models over their Eurostat-inspired counterparts for both the retail and industry series, though a substantial number of plots did not favor either approach.

Additionally, when analyzing the U.S. series, the subjective nature of the Eurostat regressors is interesting in that it is impossible to know for sure which holidays should be

accounted for and whether or not they should be treated as Sundays. In general, the scope of the holiday correction made by Eurostat is to facilitate comparisons between EU contries. While it is certainly possible that properly defined holiday correction regressors could outperform traditional trading day regressors for individual series, the subjective and individualistic nature of developing such constructions would be prohibitive. Thus, a reasonable conclusion that can be drawn from our analysis of the 93 U.S. monthly economic flow series is that there is no indication that the method of providing a country specific holiday correction to trading day regressors, by way of treating holidays as Sundays, is capable of outperforming the current X-12-ARIMA method in adjusting for trading day and holiday effects across a wide range of U.S. monthly economic flow series. However, it does appear that the Eurostat-inspired method of adjusting for trading day is capable of outperforming the current U.S. Census Bureau method for a limited number of U.S. series, and the results presented here clearly warrant further investigation. In particular, it is of interest to know what leads one model to be preferred over the other. The study of the theoretical aspects surrounding this question and methods for improving the X-12 ARIMA and Eurostat trading day adjustment methods for fixed holiday effects is currently an avenue for future research.

6. References

Bell, W.R. (2004). On RegComponent Time Series Models and Their Applications. In State Space and Unobserved Component Models: Theory and Applications, A. Harvey, S. Koopman, and N. Shephard (eds). Cambridge: Cambridge University Press, 248–283.

Burnham, K.P. and Anderson, D.R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach, (Second Edition). New York: Springer.

Burnham, K.P. and Anderson, D.R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological Methods and Research, 33, 261–304.

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.

Eurostat European Commission (2009). ESS Guidelines on Seasonal Adjustment, Luxemburg: Office for Official Publications of the European Communities, ISBN 978-92-79-12307-8.

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. Journal of Business and Economic Statistics, 16, 127–177.

Findley, D.F. and Soukup, R.J. (2000). Modeling and Model Selection for Moving Holidays. U.S. Census Bureau, Statistical Research Division.

Gómez, V. and Maravall, A. (1996). Programs TRAMO and SEATS, Instructions for the Use. Banco de Espana – Servicio de Estudios, Documento de Trabajo no. 9628 (English version).

Hurvich, C.M. and Tsai, C.L. (1989). Regression and Time Series Model Selection in Small Samples. Biometrika, 76, 297–307.

McElroy, T. and Holan, S. (2009). A Nonparametric Test for Residual Seasonality. Survey Methodology, 35, 67–83.

Otto, M.C., Bell, W.R., and Burman, J.P. (1987). An Iterative GLS Approach to Maximum Likelihood Estimation of Regression Models With ARIMA Errors. American Statistical Association, Proceedings of the Business and Economics Statistics Section, 632–637.

Shumway, R.J. and Stoffer, D.S. (2006). Time Series Analysis and Its Applications: With R Examples (Second Edition). New York: Springer.

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. U.S. Census Bureau, Statistical Research Division.

Soukup, R.J. and Findley, D.F. (2000). Detection and Modeling of Trading Day Effects. U.S. Census Bureau, Statistical Research Division.

Statistical Office of the European Communities (2002). Demetra 2.0 User Manual (Release Version 2.0).

U.S. Census Bureau (2007). X-12-ARIMA Reference Manual (Version 0.3), Washington, DC.

U.S. Census Bureau (2008). X-13A-S Reference Manual (Version 0.1 Beta), Washington, DC.

Received July 2009
Revised December 2009