Statistica Sinica **26** (2016), 1365-1388
doi:http://dx.doi.org/10.5705/ss.2014.189t

REG-ARIMA MODEL IDENTIFICATION: EMPIRICAL EVIDENCE

Agustín Maravall[1], Roberto López Pavón[2] and Domingo Pérez Cañete[2]

[1]*Bank of Spain and* [2]*Indra*

Abstract: The results of applying the default Automatic Model Identification of program TRAMO to a set of 15,642 socio-economic monthly series are analyzed. The series cover a wide variety of activities and indicators for a large number of countries, and the number of observations ranges between 60 and 600. The model considered by the automatic procedure is an ARIMA model with -when detected- outliers and calendar effects. For series with no more than 360 observations the results are found satisfactory for slightly more than 90% of the series, excellent indeed as far as whitening of the series and the capture of seasonality are concerned. For longer series the normality assumption is the weak point. Still, in so far as kurtosis is the main cause, non-normality does not seem to be a dramatic feature. The relevance of including possible outliers and calendar effects is discussed in an Appendix.

Key words and phrases: Automatic model identification, outliers, programs TRAMO and SEATS, regression-ARIMA models, seasonality, time series analysis.

1. Introduction

An issue of applied relevance in short-term economic monitoring and policy making is the error in seasonally adjusted (SA) series. Lacking a precise definition of the (unobserved) seasonality they are attempting to capture, standard methods based on (perhaps a limited set) of fixed filters cannot yield any light. A modeling approach to seasonal adjustment could solve the problem; of course, the model should be in agreement with the observations. Seasonal adjustment is routinely performed on very many series and the dynamic structure of these series, in general, differ. For a model approach to be feasible a reliable and efficient automatic model identification (AMI) procedure is needed. Based on prior work of Hillmer and Tiao (1982) and Burman (1980), Gómez and Maravall developed a pair of programs, TRAMO and SEATS, where TRAMO identifies the model for the observed series, which is then seasonally adjusted by SEATS. The paper analyzes the performance of the AMI contained in program TRAMO when applied to a set of close to 16,000 monthly series. For series comprising at most 30 years of observations the results are satisfactory, in particular in terms of the detection of unit roots and presence/absence of seasonality in the series, whitening of the series, and idempotency properties. The results of a battery of

tests (that includes tests for normality and short-term out-of-sample forecasting) show that, for 90.3% of the series, AMI yields an acceptable model.

2. Model-based Seasonal Adjustment: The Need for an Automatic Model Identification Procedure

Short-term monitoring and economic policy are mostly based on the evolution of seasonally adjusted (SA) series. Being unobserved, an estimator of the SA series is used, often obtained with an X11-X12-type filter. The estimator inevitably contains an estimation error that may well induce under and over-estimation of the SA series which, in turn, can cause policy errors. To know the distribution of this error, and in particular its standard error (SE), would be of help. More than half a century ago, Morgenstern expressed his conviction that the single step that would most contribute to transform Economics into a serious discipline would be to always present the data with the associated SE. In 1976 and 1980 the reports -to the US Federal Reserve Board- of the Bach and Moore Committees stressed the need to know the SE of the SA data (Bach et al. (1976), Moore et al. (1981)). An example of the practical importance of errors in SA data is the following. Maravall and Pierce (1986) looked at US monetary policy in the 70s, based in essence on setting an annual target for the growth of M1 and monthly ranges for the intrayear annualized monthly growth of the SA series obtained with X11. If actual growth exceeded the upper limit, Federal Funds had to go up; if it fell short of the lower limit, the Funds had to go down; otherwise, they should be left untouched. Because X11 is a two-sided symmetric filter centered on the present month, control had to use the concurrent estimator of the SA series, that is, the estimator for the last observed period, and obtained with a one-sided filter. As new observations became available and the 2-sided filter approached completion, the estimator of SA M1 was revised until it became the final or 'historical' estimator. The difference between the historical and the preliminary estimator represents an error in the latter; it is denoted 'revision error.' Maravall and Pierce compared the two estimators and computed the frequency of disagreement in terms of policy action: for close to 40% of the months, the historical estimator would have implied a different Fed reaction. The width of the (ad-hoc-set) ranges would seem inadequate, and knowledge of the SE of the revision error would certainly have been of help. Unfortunately, the lack of an underlying model for the X11-X12 methods prevented this.

As Hawkins and Mlodinow (2010) state, 'there can be no model-independent test of reality'; a way to solve the problem would be to specify a model for the observed data, from which a model for the (unobservable) seasonal component can be derived. The check for whether the SA series estimator is in agreement with the theoretical model would provide the basic tools for diagnostics and inference and, in particular, the SE of the estimation error would be easily obtained.

Hillmer and Tiao (1982) and Burman (1980) proposed a seasonal adjustment method based on ARIMA models, that came to be known as the 'Arima-model-based' (AMB) approach. The method identifies an ARIMA model for the observed series, and a partial fractions decomposition of its spectrum yields the spectra of the unobserved components that aggregate into the series. These are the trend, seasonal, and transitory/irregular components. (The term spectrum also denotes the pseudo-spectrum of a non-stationary series.) Then, the minimum-mean-square-error (MMSE) estimator of the SA series can be obtained. From the spectral decomposition, the underlying ARIMA model for each component is straightforwardly derived.

The AMB approach was attractive. Consistency between the unobserved component models and the model for the observed series implied that, if the latter is well specified - a testable assumption- spurious results (such as removing seasonality from a non-seasonal series), or cases of over/under-estimation of seasonality, would be avoided. Further, the model-based structure could be exploited to build diagnostics and inference so that, for example, the SE of the SA series and of the forecasts thereof, as well as the SE and speed of convergence of the future revision in preliminary estimators, could be obtained. However, the approach faced drawbacks. First, it seemed to require heavy use of time series analysts and computational resources. Second, many time series need pre-adjustment before ARIMA models can be assumed. For example, the series may contain outliers, calendar effects, missing observations, and be affected by intervention/regression effects. As a consequence the AMB method was not considered viable for routine and large-scale use, and hence for official data production. (This was in fact the conclusion of the Moore Committee.) Most notably, viable large-scale application requires a reliable and efficient automatic model identification (AMI) procedure.

In the mid 90s, Gómez and Maravall (1996) produced the first version of two linked programs: TRAMO ('Time series Regression with ARIMA noise, Missing values and Outliers') and SEATS ('Signal Extraction in ARIMA Time Series'). The programs contained a complete model-based application that included an AMI procedure. The model is an ARIMA model extended to include pre-adjustment through regression. This extended model is referred to as a reg-ARIMA model. In what follows, the Windows version of the two linked programs - program TSW- will be used; it can be freely downloaded from the Bank of Spain website (`www.bde.es` → Services → Statistical and Econometrics software).

Automatic default application of TSW performs outlier detection and correction, identification and ML estimation of the reg-ARIMA model, interpolation of missing values, whitening of the series, MMSE forecasting, calendar adjustment, and MMSE estimation and forecasting of the seasonal, trend, cycle, and transitory/irregular components (the SE of all estimators and forecasts are also provided).

The programs are presently used throughout the world by many agencies, institutions, firms, and universities. Possibly, the most frequent application is seasonal adjustment, where many thousands of series may have to be routinely adjusted. The programs have been widely recommended (see, for example, EUROSTAT (2009), and United Nations (2011) and, together with X12ARIMA (Findley et al. (1998)), they are part of the new X13-ARIMA-SEATS program of the U.S. Bureau of the Census and of the program JDEMETRA+ of the European Statistical System (see the two reference manuals in Wikipedia).

The present paper discusses the performance of the TRAMO-SEATS AMI procedure on a set of close to 16,000 real-world-socio-economic time series. To assess the reliability of the procedure when all parameters are set at their default values, two types of failure need to be addressed. First, for a series that follows a reg-ARIMA model, is the proper model obtained? Second, is the reg-ARIMA specification appropriate for modeling real series? The first question was addressed in Maravall, López, and Pérez (2015) and the AMI procedure was found highly reliable when a reg-ARIMA specification is appropriate.

In this paper the second question is addressed: are real series properly modeled with the reg-ARIMA specification of the default AMI procedure?

3. Summary of the Automatic Model Identification Procedure

3.1. The regression-ARIMA model

Let the observed time series be $z = (z_{t1}, z_{t2}, \dots, z_{tN})$ where $1 = t_1 < t_2 < \cdots < t_N = T$. (There may be missing observations and the series may have been log transformed.) The Reg-ARIMA model can be expressed as

$$z_t = y_t^{'}\beta + x_t, \tag{3.1}$$

where y_t is a matrix with n regression variables, β is the vector of the regression coefficients and the variable x_t follows a (possibly nonstationary) ARIMA model. Hence $y_t'\beta$ represents the deterministic component, and x_t the stochastic one. If B denotes the backward shift operator, such that $B^j z_t = z_{t-j}$, the ARIMA model for x_t is of the type

$$v_t = \delta(B)x_t, \tag{3.2}$$

$$\phi(B)[v_t - \mu_v] = \theta(B)a_t, \qquad a_t \sim niid(0, V_a), \tag{3.3}$$

where v_t is the stationary transformation of x_t, μ_v the mean of v_t, and $\delta(B)$ contains regular and seasonal differences; $\phi(B)$ is a stationary autoregressive (AR) polynomial in B and $\theta(B)$ is an invertible moving average (MA) polynomial in B. For seasonal series, the polynomials typically have a 'multiplicative' structure. Letting s denote the number of observations per year, in TRAMO, the polynomials in B factorize as

$$\delta(B) = (1-B)^d(1-B^s)^{d_s} = \nabla^d\nabla_s^{d_s},$$

where $\nabla$and $\nabla_s$are the regular and seasonal differences, and

$$\phi(B) = \phi_{p(B)}\Phi_{p_s}(B^s) = (1 + \phi_1 B + \cdots + \phi_p B^p)(1 + \phi_S B^s), \tag{3.4}$$

$$\theta(B) = \theta_q(B)\Theta_{q_s}(B^s) = (1 + \theta_1 B + \cdots + \theta_q B^q)(1 + \theta_s B^s). \tag{3.5}$$

In what follows, the variable v_t will be assumed centred around its mean and the general expression for the model will be the ARIMA $(p, d, q)(p_s, d_s, q_s)_s$ model:

$$\phi_p(B)\Phi_{p_s}(B^s)\nabla^d\nabla_s^{d_s}x_t = \theta_q(B)\Theta_{q_s}(B^s)a_t, \tag{3.6}$$

where the orders considered by AMI are restricted to $p, q = 0, 1, 2, 3$; $d = 0, 1, 2$; $d_s, p_s, q_s = 0, 1$.

In what follows, the only regression variables are the outliers and Calendar effects automatically identified by the default run of the program. Three types of possible outliers are considered: additive outlier (AO), a single spike; transitory change (TC), a spike that takes several periods to return to the previous level; and level shift (LS), a step function. Calendar effects are Trading Day (with a day-of-week or a working/non-working day specification), Easter, and Leap Year. TRAMO pre-tests for the log/level transformation, and performs automatic ARIMA model identification joint with automatic outlier and calendar effect detection, estimates the model by exact maximum likelihood, interpolates missing values, and forecasts the series.

3.2. Automatic model identification in the presence of outliers

The algorithm iterates between two stages.

1. Automatic outlier detection and correction. The procedure is based on Tsay (1986) and Chen and Liu (1993) with some modifications (Gómez and Maravall, 2001a). At each stage, given the ARIMA model, outliers are detected one by one, and eventually jointly estimated, together with calendar effects (if present) by GLS.
2. Automatic model identification. TRAMO proceeds by iterating two steps: First, it identifies the differencing polynomial $\delta(B)$ that contains the unit roots. Second, it identifies the ARMA model, $\phi_p(B), \Phi_{p_s}(B^s), \theta_q(B)$, and $\Theta_{q_s}(B^s)$. A pre-test for possible presence of seasonality determines the default model, used at the beginning of AMI and at some intermediate stages (as a benchmark comparison). For seasonal series the default model is the so-called Airline model (Box and Jenkins (1970)), given by

$$\nabla\nabla_s x_t = (1 + \theta_1 B)(1 + \theta_S B^s)a_t, \tag{3.7}$$

the IMA $(0, 1, 1)(0, 1, 1)_s$ model. For nonseasonal series the default model is

$$\nabla x_t = (1 + \theta B)a_t + \mu, \tag{3.8}$$

the IMA (1,1) plus mean model. Identification of the ARIMA model is performed with the series corrected for Calendar effects and for the outliers detected at that stage. If the model changes, the automatic detection and correction of outliers is performed again from the beginning. Intermediate stages employ the Hannan and Rissanen (1982)method. Final estimation always uses the exact maximum likelihood method of Gómez and Maravall (1994).

3.3. Identification of the nonstationary polynomial $\delta(B)$

To determine the appropriate differencing of the series standard unit root tests are discarded. First, when MA roots are not negligeable, the standard tests have low power. Second, a run of AMI for a single series may try thousands of models, where the next try depends on previous results. There is, thus, a serious data mining problem: the size of the test is a function of prior rejections and acceptances, and its correct value is not known.

We follow an alternative approach that relies on the superconsistency results of Tiao and Tsay (1983), and Tsay (1984). Sequences of multiplicative AR(1) and ARMA(1,1) are estimated, and instead of a fictitious size, the following value is fixed a priori: how large should the modulus of an AR root be in order to accept it as 1? By default, in the sequence of AR(1) and ARMA(1,1) estimations, when the modulus of the AR parameter is above 0.91 (seasonal polynomial) or 0.96 (regular polynomial), it is made 1. Unit AR roots are identified one by one; for MA roots invertibility is strictly imposed and the maximum allowed modulus is 0.95.

3.4. Identification of the stationary ARMA model: $\phi(B)$ and $\theta(B)$

Identification of the stationary part of the model attempts to minimize the Bayesian information criterion

$$BIC_{P,Q} = \ln(\hat{\sigma}^2_{P,Q}) + (P+Q)\frac{\ln(N-D)}{N-D},$$

where $P = p+p_s$, $Q = q+q_s$, and $D = d+d_s$. The search is done sequentially: for fixed regular polynomials, the seasonal ones are obtained, and viceversa. A more complete description of the AMI procedure and of the estimation algorithms can be found in Gómez and Maravall (1994, 2001a); Gómez, Maravall, and Peña (1999); and Maravall and Pérez (2012). Program SEATS is described in Gómez and Maravall (2001b).

4. Application to a Large Set of Series: Empirical Results

The default automatic option of program TSW was applied to 15,642 monthly series obtained from a variety of sources over a period of 30 years. The geographical distribution is the following: 14% are Spanish series, 42% are from other European countries, 24% are US series, 20% cover the rest of the world. All sorts of

Table 1. General.

Group (by NZ)	# series in group	Average length	% logs	Average # parameters/ser	Average # outliers/ser	% with calendar effect	% with outliers
60-110	3972	89	83.4	2	0.8	28.5	43.2
111-160	4652	130	91.9	2.1	1.5	62.1	63.7
161-210	3065	173	91.4	2.3	1.7	67.6	67.8
211-260	2101	229	82.1	2.2	2.5	28.2	82
261-360	1009	290	84.4	2.5	2.9	57.4	86.5
TOTAL	**14799**	**153**	**87.6**	**2.2**	**1.6**	**49.2**	**63.2**
361-600	843	456	77.1	2.9	4	36.3	74

economic activities and indicators are represented. The number of observations (NZ) ranges between 60 and 600 observations (this is the range covered by the standard TSW AMI). The series have been grouped according to length, and the number of series in each group intends to roughly reflect the relative frequency of the lengths encountered in practice. The content of the groups is fairly heterogeneous. Table 1 presents some general results (NZ denotes the length of the series.)

For reasons given later, when computing total averages only the interval (60−360 observations) is considered. The longer-series group was relatively small and, for short-term forecasting and seasonal adjustment, the adequacy of the reg-ARIMA specification to series exceeding 360 monthly observations is more questionable. (Long series were split into two and added to the shorter sample size groups.)

Table 1 shows that for most of the series logs are chosen, that the average number of parameters lies in the range 2 - 2.9 parameters per series and increases moderately with length, that for all groups outliers are found at a rate of approximately 1 outlier per 100 observations, that slightly less than half of the series require calendar adjustment, and that about two thirds require outlier adjustment. (Many series, of course, share outliers.) Table 2 shows that most of the outliers are additive outliers, followed by level shifts. Table 3 indicates that the preferred Trading Day specification is the parsimonious working/non-working days. The Easter effect is detected in less than 16% of the series.

5. Presence of Seasonality

The model that starts AMI depends on whether seasonality has been detected or not in the series. The detection is based on four separate checks: a χ^2_{11} non-parametric rank test similar to the one in Kendall and Ord (1990); a check of the autocorrelations for seasonal lags (12 and 24) in the line of Pierce (1978),

Table 2. Outliers.

Group (by NZ)	Average # per series			
	AO	TC	LS	Tot.
60-110	0.4	0.2	0.3	0.8
111-160	0.7	0.3	0.5	1.5
161-210	0.8	0.4	0.5	1.6
211-260	1.1	0.5	0.9	2.5
261-360	1.6	0.6	0.7	2.9
TOTAL AVERAGE	**0.8**	**0.3**	**0.5**	**1.6**
361-600	1.8	1.0	1.3	4.0

Table 3. Calendar effects.

Group (by NZ)	% of series with				
	TD		EE	STOCH. TD	Total calendar effect
	1 var	6 var			
60-110	18.4	7.3	6.7	0.3	28.5
111-160	55.2	3.7	17.2	1.9	62.2
161-210	59.9	6.2	24.7	2.6	67.8
211-260	19.7	6.6	11.9	1.4	28.2
261-360	48	5.5	28.4	2	57.4
TOTAL AVERAGE	**40.8**	**5.7**	**15.9**	**1.6**	**49.2**
361-600	25.4	8.4	12.5	3.7	36.3

using a χ^2_2; an F-test for the significance of seasonal dummy variables similar to the one in Lytras, Feldpausch, and Bell (2007); a test for the presence of peaks at seasonal frequencies in the spectrum of the first differenced series. The first three tests are applied at the 99% critical value. The fourth test combines the results of two spectrum estimates: one obtained with an AR(30) fit in the spirit of X12-ARIMA Findley and Martin (2006); the second is a non-parametric Tukey-type estimator, as in Jenkins and Watts (1968), that we approximate by an F distribution. The results of the four tests are combined into an 'overall' test that answers the question: can seasonality be assumed to be present? The tests are first applied to the original series and determine the starting model in AMI. Once the series has been corrected for regression effects (outliers and calendar effect), the tests are applied again to the 'linearized' series; these are the results reported in Table 4.

We consider that only for series with 80 or more observations are the spectra worth estimating. Thus the 45.3% in Table 4 is a strongly downward-biased estimator.

Table 4. Detection of seasonality.

Group	Pre-tests: % of series with seasonality					Seasonal component in AMI model	Warning-free SA series
	NP	QS	Spectrum	F	Overall test		
60-110	81.1	82.8	45.3	83.9	85.6	85.5	81.9
111-160	77.8	75.6	76.0	79.7	80.6	79.8	76.6
161-210	85.7	84.3	82.6	88.5	89.7	88.8	85.3
211-260	83.2	81.6	82.8	85.9	86.8	85.5	82.7
261-360	82.7	80.7	81.2	86.3	88.2	85.8	84.4
TOTAL AVERAGE	**81.4**	**81.0**	**70.4**	**84.0**	**85.2**	**84.4**	**81.7**
361-600	75.9	75.0	75.1	74.9	77.9	76.8	75.3

Presence/absence of seasonality may again be tested at intermediate steps of AMI. Seasonality is assumed even when the evidence is weak, so that the overall test favors overdetection. The AMI procedure itself corrects cases of seasonality overdetection, and this occurs in 1% of the series for which seasonality had been detected. Further, SEATS may detect cases in which seasonality is not well-behaved (e.g., excessively erratic) and its MMSE estimator of little use (e.g., when the revision error is excessively high). SEATS produces then a warning questioning the quality of the adjustment; this affects 3.2% of the 13,525 series for which AMI produced a model with seasonality.

The presence-of-seasonality test is again employed at the diagnostic stage, to check model residuals, SA series, Trend-cycle, and irregular components.

6. ARIMA Model

Table 5 presents the differencing needed in order to achieve stationarity of the series, and shows that the proportion of stationary series decreases monotonically as NZ increases. For the group with the shortest series 28% of them are stationary; for the group with the longest series, the proportion is 0.5%. Within the series in the range ($60 \leq NZ \leq 360$), about 10% of them are stationary, 82% of them require $D = 1$, 72% require $DS = 1$, and only 2% require $D = 2$ (11% for the series with $NZ > 360$). Most of the series (65%) require the $(D = 1, DS = 1)$ transformation.

Concerning the parameters of the stationary ARMA part of the model, Table 6 presents their average number for the regular and seasonal AR and MA polynomials. It is seen that moving average parameters are more frequent than autoregressive ones. This is due to the predominance of the Airline Model specification, which -as seen in Table 7- applies to 40% of the series. The remaining

Table 5. Differences.

Group	Stationary (no diff.)	% of series with				
		∇	∇^2	∇_{12}	$\nabla\nabla_{12}$	$\nabla^2\nabla_{12}$
60-110	28.1	16.8	0.2	7.9	45.1	1.9
111-160	5	19.2	1.9	6.5	66.9	0.4
161-210	3.4	13.5	0.2	4.3	78	0.5
211-260	2.2	17.5	1	3.7	73.4	2
261-360	1.8	15	0.7	2.8	78.3	1.5
TOTAL AVERAGE	**10.3**	**16.8**	**0.9**	**5.8**	**65**	**1.1**
361-600	0.5	20.4	6.2	4.7	63.5	4.7

Table 6. ARMA parameters.

Group	Average # per series				
	P	Q	BP	BQ	Total
60-110	0.6	0.6	0.3	0.5	2
111-160	0.5	0.8	0.1	0.8	2.1
161-210	0.5	0.9	0.1	0.9	2.3
211-260	0.5	0.7	0.1	0.8	2.2
261-360	0.5	1	0.1	0.9	2.5
TOTAL AVERAGE	**0.5**	**0.8**	**0.2**	**0.7**	**2.2**
361-600	1	0.9	0.1	0.8	2.9

60% comprise 246 different ARIMA specifications (out of a total of 384 possible ones); 24 of these specifications account for 40% of series and the remaining 222 specifications account for 20% of them. Table 7 displays the 30 specifications most frequently encountered.

7. Residual Diagnostics and Out-of-sample Performance

TSW+ offers two basic types of diagnostics. One is aimed at testing the normally, identically, and independently distributed assumption on the residuals; the other performs out-of-sample forecast tests. The normality assumption is checked with the Bera-Jarque normality test, plus the skewness and kurtosis t-tests; the autocorrelation test is the standard Ljung-Box test (with 24 autocorrelations); independence is further checked with a non-parametric t-test on randomness of the residual sign runs; and the identical distribution assumption is checked with the constant zero mean and constant variance tests that compare the first and second half of the series residuals. The results of these tests are given in Table 8. The overall test for seasonality is applied to the residuals, as well as a spectral test to check for a residual Trading Day effect (Table 9). The

Table 7. Original series.
Total number of identified model orders: 246
Most frequent ones: $(p, d, q)(bp, bd, bq)_{12}$

P	D	Q	BP	BD	BQ	%
0	1	1	0	1	1	40
0	1	1	0	0	0	4
1	1	0	0	1	1	4
0	1	0	0	1	1	3
1	0	0	1	0	0	3
2	1	0	0	1	1	2
0	1	2	0	1	1	2
1	0	0	0	1	1	2
1	1	1	0	1	1	2
0	1	1	1	0	0	2
1	1	0	0	0	0	2
3	1	1	0	1	1	1
1	0	0	0	0	0	1
1	0	1	1	0	0	1
3	1	0	0	1	1	1
0	1	1	0	1	0	1
2	0	0	0	1	1	1
0	1	1	1	1	1	1
2	1	1	0	1	1	1
0	1	3	0	1	1	1
0	1	0	1	0	0	1
0	1	2	0	0	0	1
0	1	0	0	0	0	1
1	0	1	0	1	1	1
0	2	1	0	1	1	1
3	0	0	0	1	1	1
0	0	0	0	1	1	1
2	1	0	0	0	0	1
2	0	0	1	0	0	1
0	1	1	0	0	1	1

out-of-sample checks are, first, a test whereby one-period-ahead forecast errors are sequentially computed for the model estimated for the series with the last 18 observations removed (the model remains fixed), and an F-test that compares the variance of these errors with the variance of the in-sample residuals.

The second test computes the standardized out-of-sample one-period-ahead forecast error for each of the series in the group, and computes the proportion that lie beyond the 1% critical value of a t distribution. (The option TERROR,'TRAMO for errors,' applied to the full group, directly provides the an-

Table 8. Residual diagnostics.

Group	mean =0	Mean stability	Variance stability	Autocorr.	Random signs	Skewness	Kurtosis
60-110	1.1	2.7	2.7	0.4	0.5	2.5	2
111-160	0.7	2.3	5.7	1.5	0.5	3.1	5.2
161-210	0.3	0.1	3.1	1.3	0.6	3.9	6.1
211-260	0.4	0.7	2.9	2	0.7	4.5	14.4
261-360	0.4	0.4	4	2.7	1.1	4.8	16.9
TOTAL AVERAGE	**0.7**	**1.6**	**3.8**	**1.3**	**0.6**	**3.4**	**6.6**
361-600	0.9	1.3	11.9	14.1	3.1	12.1	49.5

Table 9. Seasonality and calendar residual effects.
(% of residual series in group that show evidence; 1% size test)

Group	Evidence of seasonality Overall test	Spectral evidence of TD effect in residuals
60-110	0.2	0.0
111-160	0.1	0.4
161-210	0.1	1.2
211-260	0.1	0.9
261-360	0.3	0.7
TOTAL AVERAGE	**0.1**	**0.5**
361-600	0.6	3.0

Table 10. Real series: Model diagnostics; % of series in group that fail the test; Out-of-sample forecast.

Group	F-test (18 final periods)	t-test (1 period ahead)
60-110	7.7	8.7
111-160	5.2	8.2
161-210	7.0	11.9
211-260	11.3	7.7
261-360	5.2	2.8
TOTAL AVERAGE	**7.1**	**8.4**
361-600	9.5	4.1

swer.) The results of these two tests appear in Table 10.

These last three tables present the results of 12 tests performed for 6 groups, each test carried at the 1% size. Centering on the 5 groups in the range $NZ = 60 - 360$, in practically all cases the residual mean can be accepted as 0 and stable. Likewise, the residuals can be assumed free of seasonality and of Trading Day effects, and their signs can be assumed random. Further, the residuals can

be accepted as uncorrelated: the percent of failures increases with NZ from 0.4% to a maximum of 2.7%. The variance is reasonably stable: the percent of failures is always $\leq 4\%$, except for the second group for which it is 5.7%.

The worst performing residual diagnostics is the kurtosis assumption. Although the average percent of failures is $< 6.6\%$, for series longer than 20 years the deterioration is remarkable. The damage however is relatively limited as far as point estimation is concerned; it mainly affects inference. More relevant is skewness; its failures are relatively moderate (3.5%).

Both out-of-sample forecast tests behave reasonably well. On average, the percent of failures of the F-test is 7%, while for the t-test it is 8%. The range of failures for the tests in Table 10 is $(3-11)\%$. For the group with more than 360 observations the percent of failures in the in-sample tests increases notably: for variance stability it goes from around 4 to 12%; for autocorrelation, from around 1 to 14%; for skewness, from around 3 to 12%, and the increase is spectacular for normality and kurtosis, for which the percent of failures jumps from 8 and 7% to 50% in both cases. This deterioration of the in-sample diagnostics does not affect out-of-sample forecasting. The relatively good forecasting performance is partly due to the small effect non-normality is likely to have on point estimation when the distribution is symmetric.

More detailed information on the performance of the tests is given in Appendix 1, where the histograms of the tests for the 6 group of series considered are displayed and compared to the (asymptotic) distribution used in each test. Concerning the in-sample residual tests, they broadly deteriorate as the series length increases. For the zero-mean randomness in signs and mean-stability tests, this deterioration is minor, and the same is true for the three residual seasonality tests (seasonal autocorrelation, non-parametric, and F-tests). For the overall residual autocorrelation and the skewness tests the deterioration is minor when $NZ \leq 360$; for normality, when $NZ \leq 260$, and for kurtosis when $NZ \leq 210$. (Kurtosis is the worst behaved test.) As for the two out-of-sample tests, for the one based on the 1-period-ahead error, the deterioration occurs as one moves from the longer to the shorter series. Finally, the test based on the 1 to 18-periods-ahead errors shows no clear deterioration in either direction.

8. Idempotency

Given that no seasonality should be present in a SA series, an important property of a seasonal adjustment method should be idempotency, which implies that seasonal adjustment of the SA series leaves the series unchanged. While fixed filters such as X11 cannot exhibit this property, it should characterize a model-based approach, where the adjustment depends on the dynamic structure of the series (see Gómez and Maravall (2001b), and Bell and Martin (2004)).

Table 11. Idempotency(% of failures).

Group	Seasonality in SA series	Warning-free SA series
60-110	0.37	0.22
111-160	0.23	0.19
161-210	0.29	0.25
211-260	0.38	0.19
261-360	1.10	0.90
TOTAL	**0.35**	**0.25**
361-600	1.18	0.59

To check whether the default automatic procedure of TRAMO-SEATS satisfies the idempotency property, the procedure was applied to the set of 15,624 series; seasonal adjustment was performed for the 13,138 of them for which seasonality had been detected. Then, the default automatic procedure was again applied to the set of SA series. Table 11 displays the % of SA series for which there is evidence of seasonality. Some evidence is found in about 1 out of 300 SA series; this seasonality is 'not questionable' (warning-free in SEATS) in 1 out of 435 SA series.

9. Conclusion: Validity of the Arima Model

When the number of observations is no more than 360, the results from the default automatic run are good, excellent indeed as far as whitening of the series and capture of seasonality are concerned. For longer series excess kurtosis is the weak point. Of course, for groups of problematic series, non-default parameter values can be entered in the automatic procedure. For example, if no outlier has been detected and the series residuals are non-normal, lowering the critical value in the outlier detection test may improve results. Alternatively, for very long series, removing observations at the beginning of the series is likely to help.

The tests in Tables 8, 9, and 10 address the performance of the model identified by AMI by looking at 12 tests. Final assessment of the quality of the model requires their combination. In TRAMO, the fit is 'good' when all tests are passed at the 1% size and the proportion of outliers is below 5%; it is 'acceptable' if it is not 'good', yet six of them (lack of autocorrelation, randomness in sign, mean stability, skewness, lack of residual seasonality, and out-of-sample F-test) are passed at the 1% level and all others at the 0.5% or 0.1% level; it is 'mildly poor' if it is neither good nor acceptable, yet all tests are passed at the 0.1% size. Otherwise the fit is judged 'poor'. Table 12 shows the quality of the fit in % of the series in the group. On average, when the sample size is no more than 360, more than 90% of the fits are good (77%) or acceptable (13.3%); when it exceeds 360 observations the percent goes down to slightly less than 50%. This

Table 12. Validity of ARIMA Model (in % of series in group).

Group	GOOD	ACCEPTABLE	MILDLY POOR	POOR	G + A
60-110	84.2	9.1	5.2	1.5	**93.3**
111-160	79.5	10.9	6.4	3.2	**90.4**
161-210	77.7	14.1	5.0	3.2	**91.8**
211-260	65.6	20.1	5.6	8.7	**85.7**
261-360	59.3	23.3	9.7	7.7	**82.7**
TOTAL	**77.0**	**13.3**	**5.9**	**3.8**	**90.3**
361-600	20.2	28.5	6.7	44.6	**48.7**

result justifies the earlier statement that for series with more than 30 years of monthly data, the performance of reg-ARIMA deteriorates significantly.

Appendix 1. Histograms of Tests and Asymptotic Distributions Used

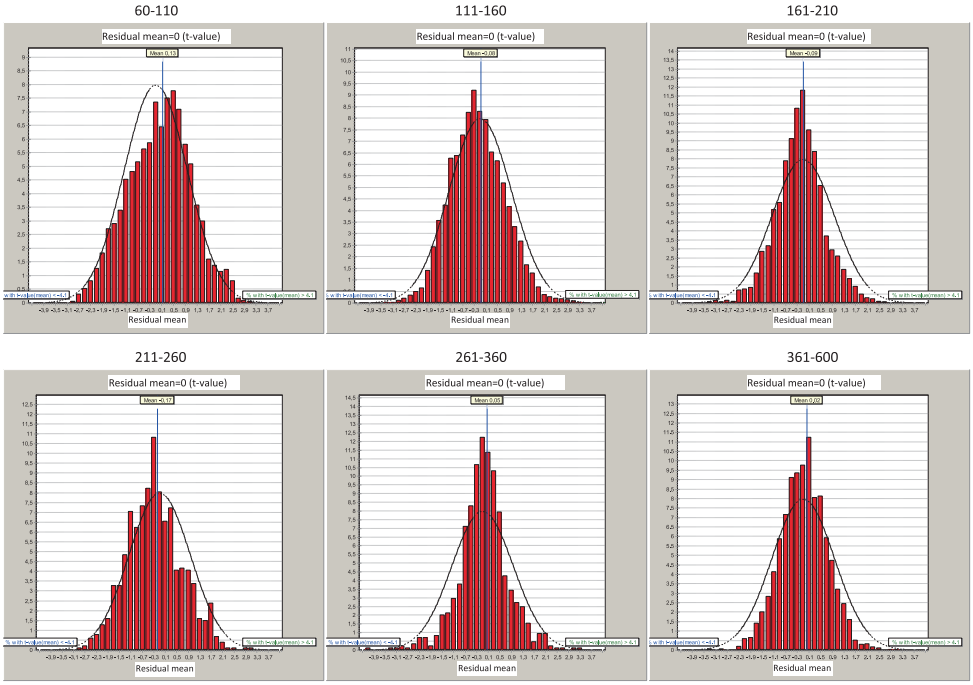

Figure 1. Zero mean of residuals test.

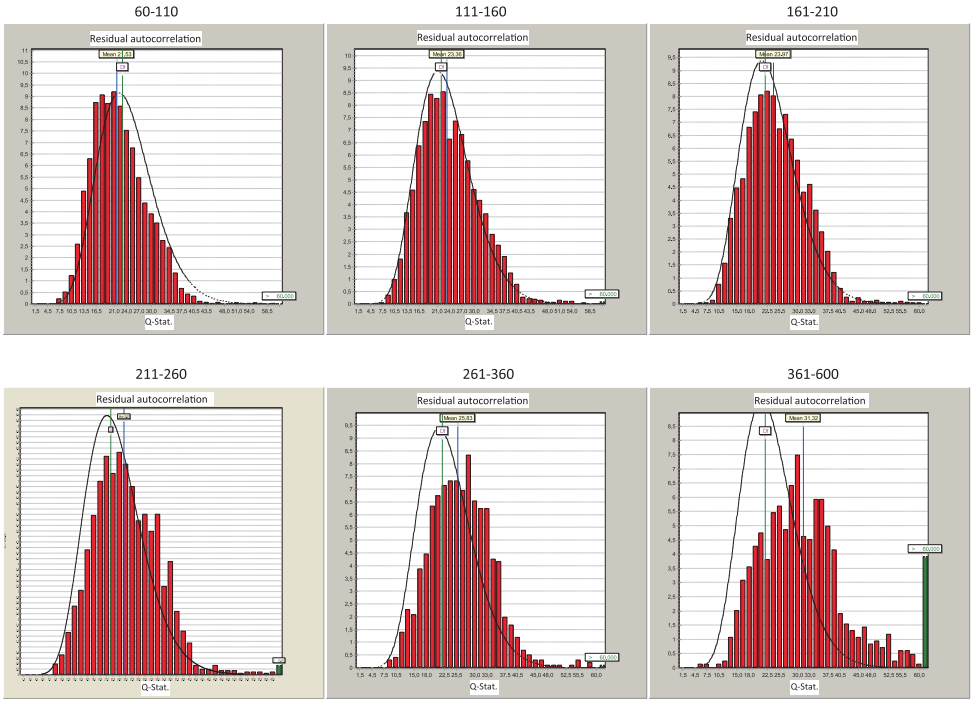

Figure 2. Residual autocorrelation test.

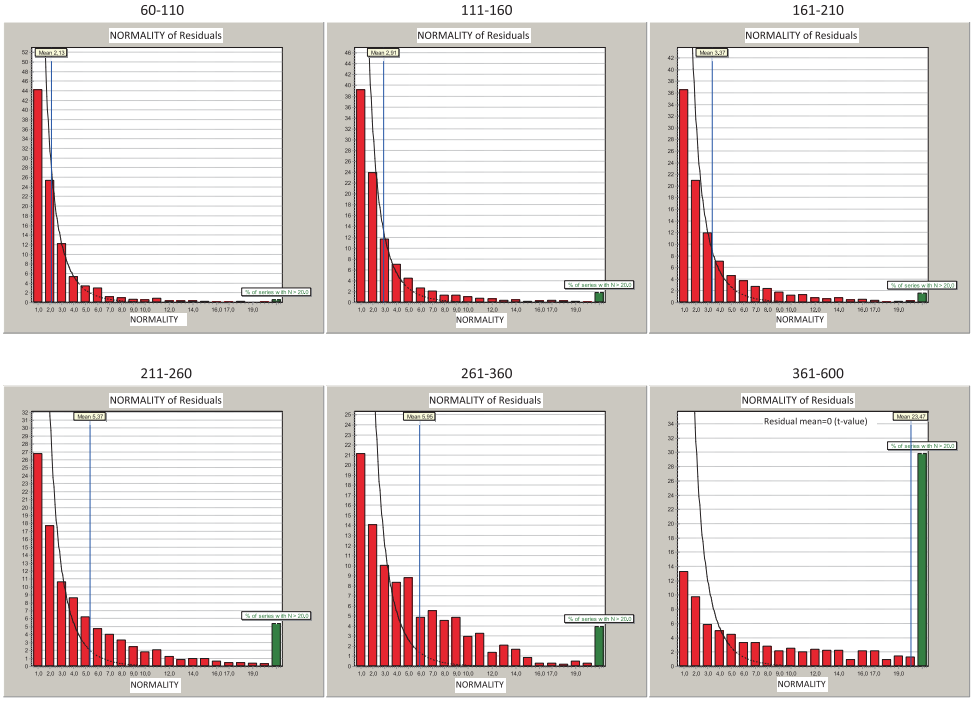

Figure 3. Normality of residuals test.

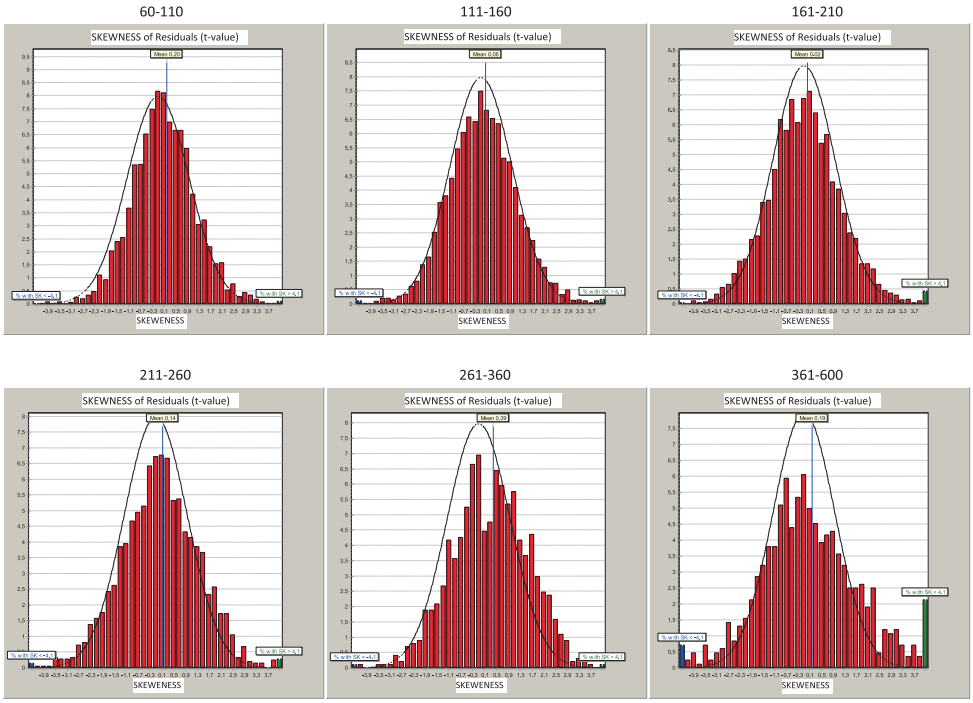

Figure 4. Skewness of residuals test.

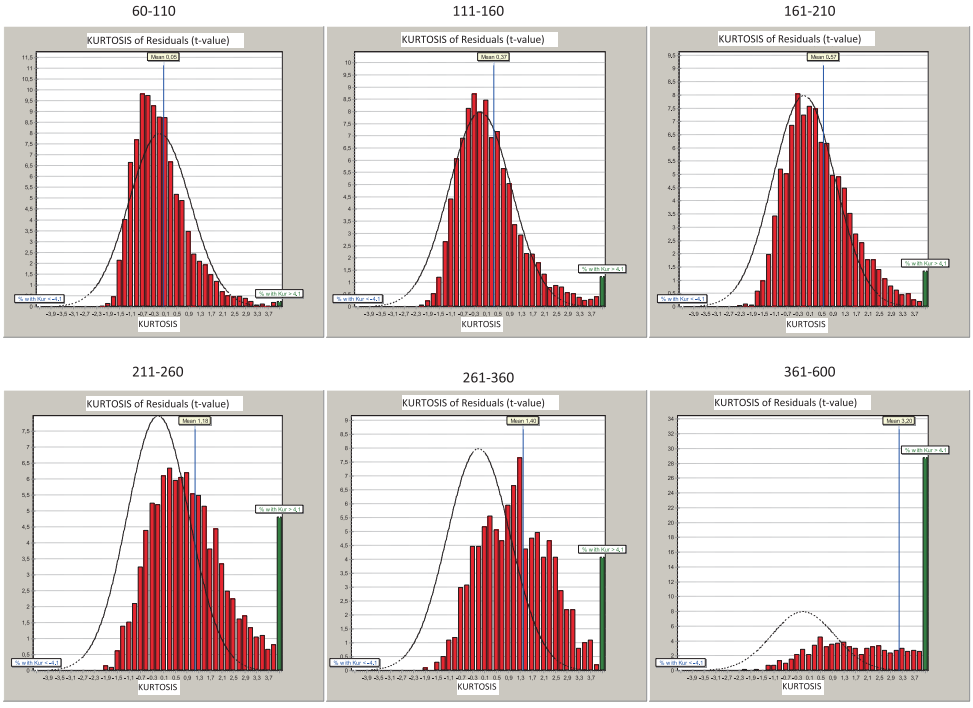

Figure 5. Kurtosis of residuals test.

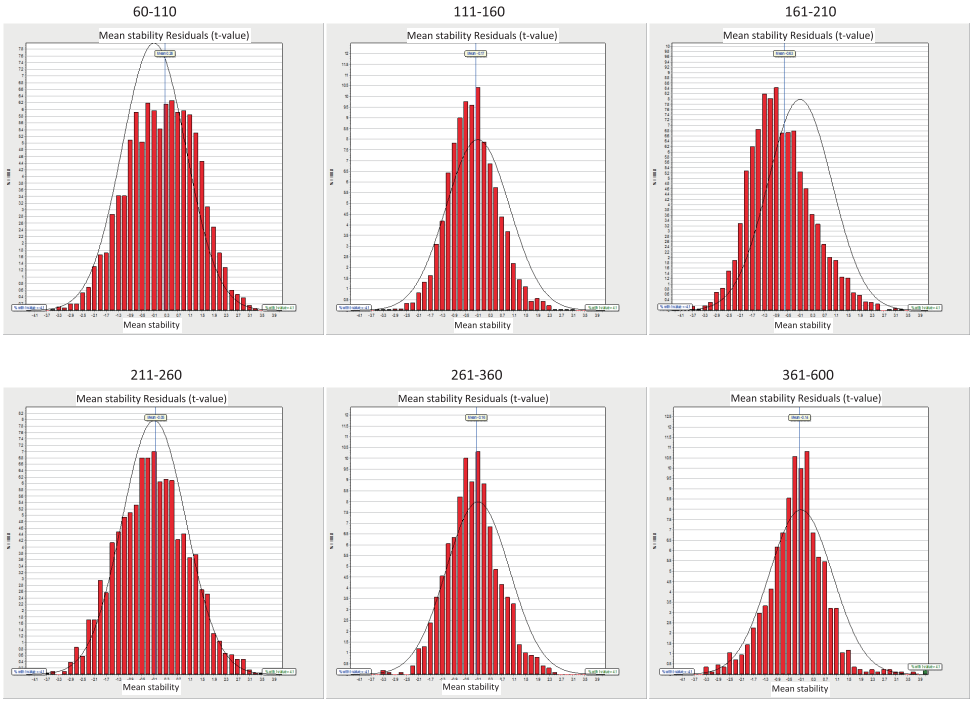

Figure 6. Mean stability test.

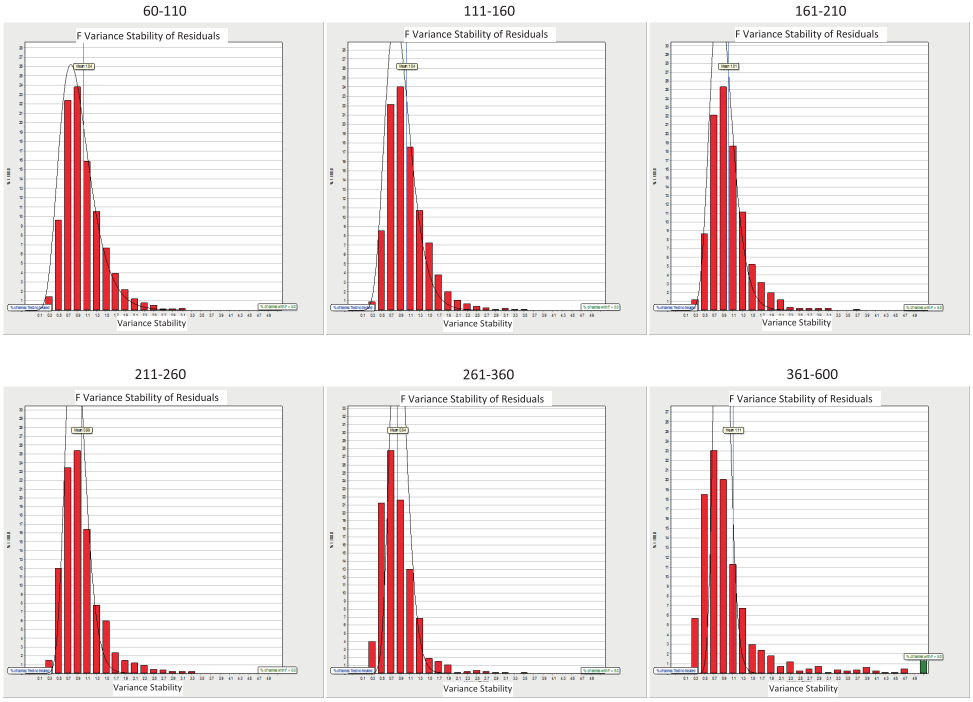

Figure 7. Variance stability test.

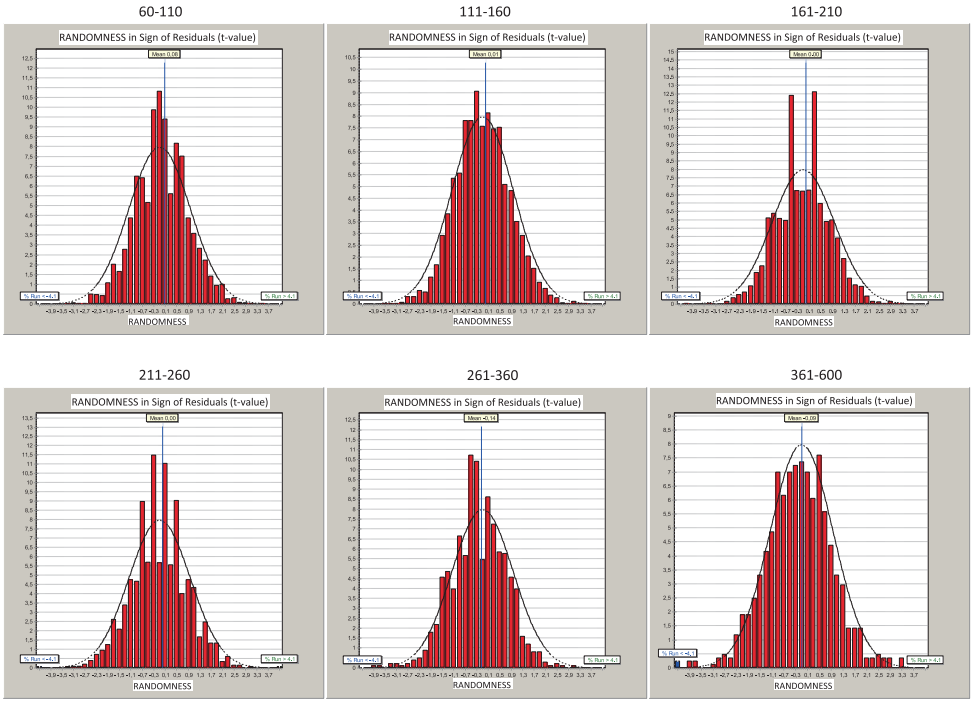

Figure 8. Randomness in signs of residuals test.

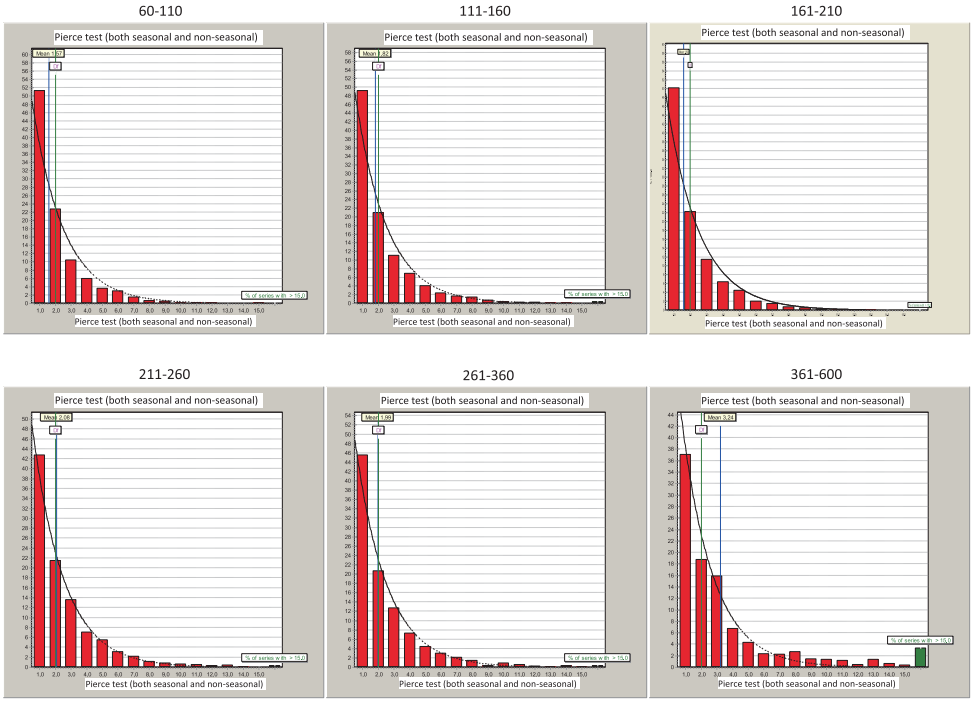

Figure 9. Seasonality: autocorrelation test.

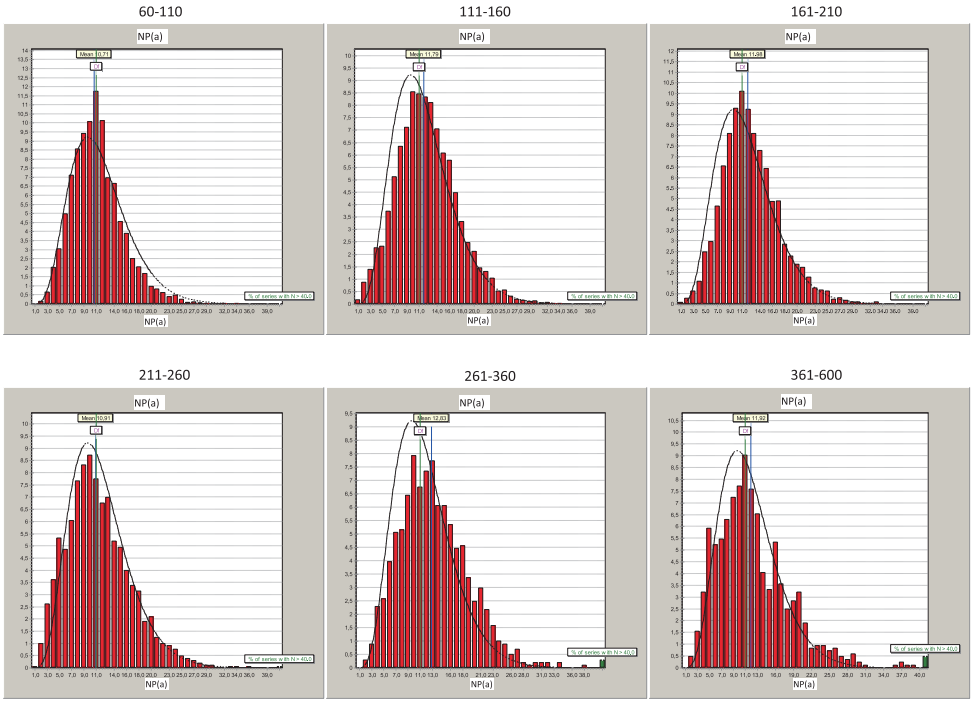

Figure 10. Seasonality: non-parametric test.

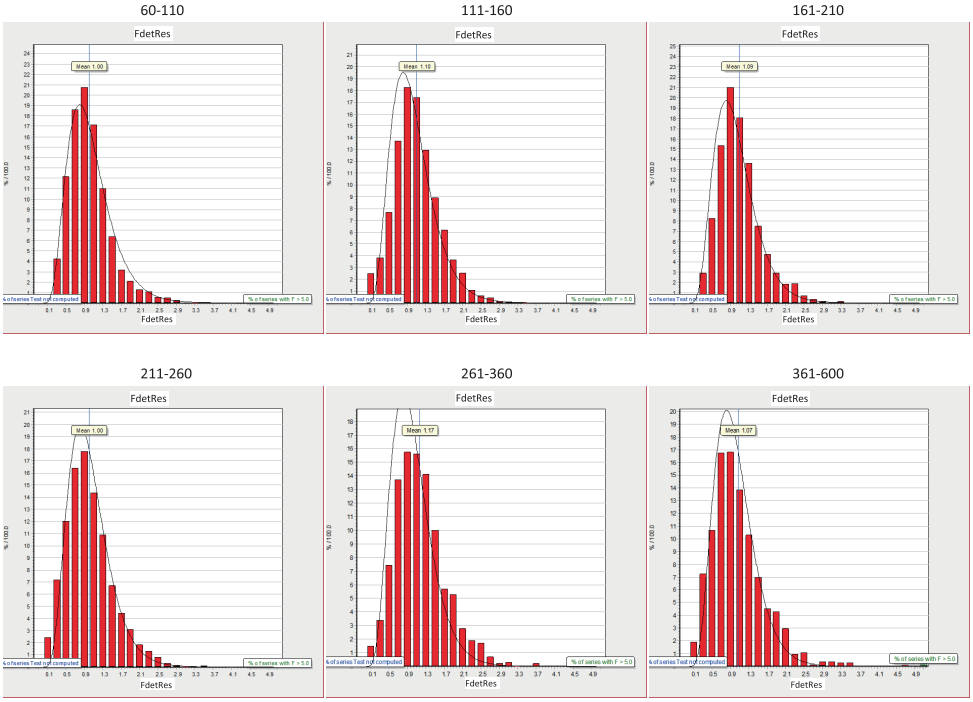

Figure 11. Seasonality: dummy variables F-test.

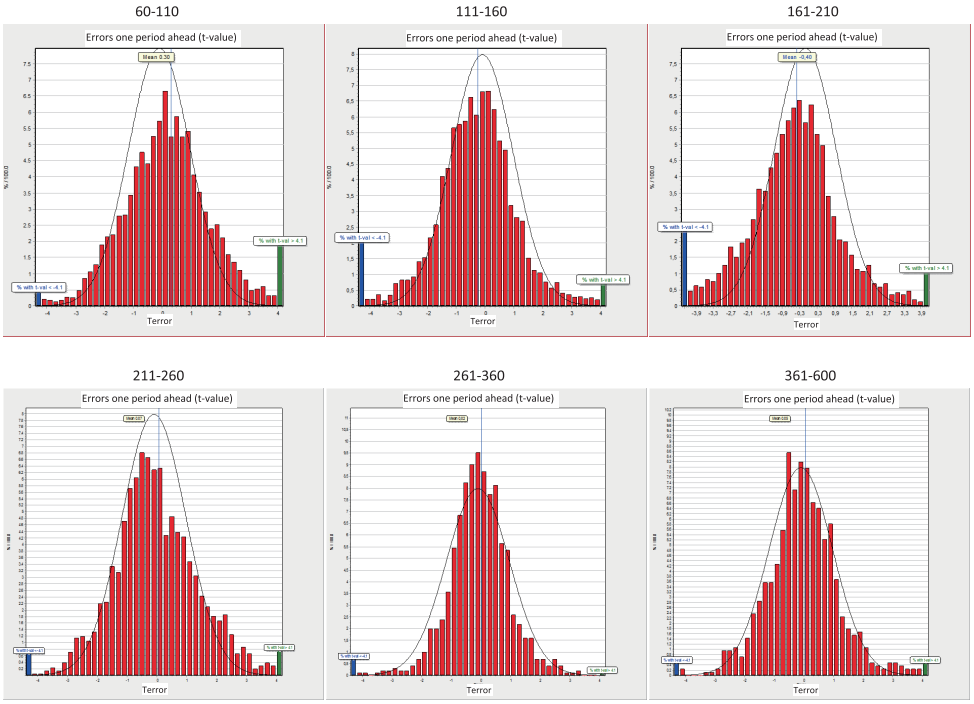

Figure 12. Out-of-sample forecast errors: 1 period-ahead test.

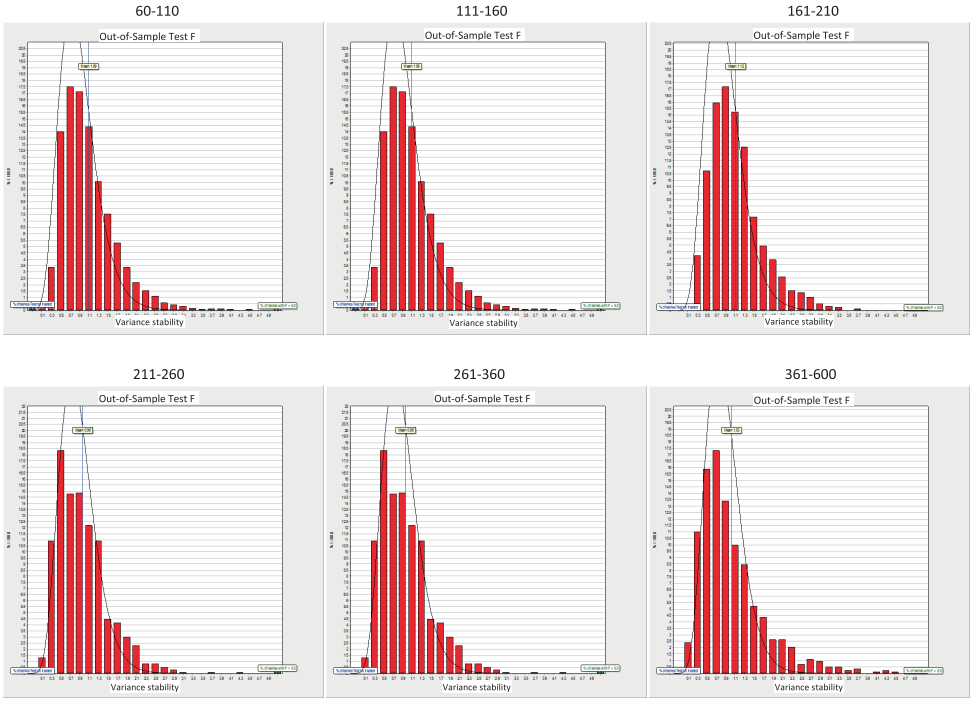

Figure 13. Out-of-sample forecast errors: 1 to 18 periods-ahead test.

Table A.1. Effect of outlier and calendar adjustment: Percent of model-fit tests failures.

Group	No autocorrelation in residuals			Normality of residuals		
	With outlier and calendar adjustment	No outlier adjustment	No outlier, no calendar adjustment	With outlier and calendar adjustment	No outlier adjustment	No outlier, no calendar adjustment
60-110	0.4	0.8	0.7	3.5	18.2	18.1
111-160	1.5	2.0	4.1	6.2	36.3	35.0
161-210	1.3	2.3	10.1	8.0	42.0	39.2
211-260	2.0	4.8	8.3	15.2	63.4	63.4
261-360	2.7	5.8	11.3	17.2	68.4	66.1
TOTAL	**1.3**	**2.4**	**5.5**	**7.9**	**38.7**	**37.5**
361-600	14.1	20.4	26.6	49.3	70.8	70.1

Note: Percents of total number of series in group.

Appendix 2. The Need for Preadjustment: Outliers and Calendar Effects

Although applied statisticians are generally aware of the need to deal with outliers and calendar effects, these effects are seldom considered in econometric applications. It is of interest to see what is their relevance in the set of series we consider.

Table A.1 shows the effect on the Ljung-Box autocorrelation and Jarque-Bera normality tests of not considering outliers, and of not considering outliers nor calendar adjustment. The effect on the autocorrelation test is relatively moderate, although ignoring outliers and calendar effects more than quadruples the % of test failures.

The effect of outliers on normality is spectacular. The 8% of failures when the two effects are considered increases to close to 40% when outliers are ignored. Calendar effects, however, do not affect normality. The two effects seem to complement each other: outliers are needed for residual normality; calendar effects help to clean residual autocorrelation. While calendar adjustment may be a convenience, outlier adjustment is a necessity.

Table A.2 presents the percent increase in the SE of the residuals when outliers and calendar effects are ignored, and both effects are seen to be significant. The improvement due to outlier removal is greater than that due to calendar adjustment, although the latter is certainly not trivial. Both effects are particularly important when the series is modelled in the levels.

The percentages in Table A.1 and A.2 have been computed for the full set of series when, in fact, only 63% of the series in the set require outlier correction, and only 50% require calendar adjustment. Thus the effect of the two corrections on a series that requires them is, on average, about 60% (outliers) and 100% (calendar effect) greater than the ones displayed in both tables. Altogether, considering

Table A.2. Effect of ignoring outlier and calendar adjustment: Percent increase in residual standard error.

Group	Series in logs		Series in levels	
	No outlier adjustment	No outlier, no calendar adjustment	No outlier adjustment	No outlier, no calendar adjustment
60-110	12.9	17.9	11.7	14.7
111-160	10.5	21.8	32.6	40.4
161-210	9.2	18.3	30.4	35.4
211-260	14.8	17.1	17.6	19.0
261-360	12.3	16.1	17.6	20.7
TOTAL	**11.6**	**19.0**	**23.4**	**28.1**
361-600	19.0	21.0	34.5	35.6

Note: First two columns: Percents of the total number of series modeled in logs in the group.
Last two columns: Id. of the series modeled in levels.

that the price paid for outlier correction is 1 outlier/100 observations, and the price paid for Calendar adjustment for the vast majority of series is the addition of 1 - perhaps 2- parameters to the model, both corrections seem worth considering.

References

Bach, G. L., Cagan, P.D., Friedman, M., Hildreth, C. G., Modigliani, F., Okun, A. (1976). *Improving the Monetary Aggregates: Report of the Advisory Committee on Monetary Statistics*. Board of Governors of the Federal Reserve System.

Bell, W. R. and Martin, D. E. (2004). Computation of asymmetric signal extraction filters and mean squared error for ARIMA component models. *J. Time Series Anal.* **25**, 603-603.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

Burman, J. P. (1980). Seasonal adjustment by signal extraction. *J. Roy. Statist. Soc. Ser. A* **152**, 321-337.

Chen, C. and Liu, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series. *J. Amer. Statist. Assoc.* **88**, 284-297.

EUROSTAT (2009). *European Statistical System Guidelines on Seasonal Adjustment*. Luxembourg: Office for Official Publications of the European Communities.

Findley, D. F. and Martin, D. E. (2006). Frequency domain analyses of SEATS and X-11/12-ARIMA seasonal adjustment filters for short and moderate-length time series. *J. of Official Statist.* **22**, 1.

Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C. and Chen, B.-C. (1998). New capabilities and methods of the X-12-ARIMA seasonal-adjustment program (with discussion). *J. Business Econom. Statist.* **12**, 127-177.

Gómez, V. and Maravall, A. (1994). Estimation, prediction, and interpolation for nonstationary sseries with the Kalman filter. *J. Amer. Statist. Assoc.* **89**, 611-624.

Gómez, V. and Maravall, A. (1996). Programs TRAMO and SEATS. Instructions for the User (with some updates). Working paper 9628. Servicio de Estudios, Banco de España.

Gómez, V. and Maravall, A. (2001a). Automatic modeling methods for univariate series. In *A Course in Time Series Analysis* (Edited by D. Peña, G. C. Tiao and R. S. Tsay), 171-201.

Gómez, V. and Maravall, A. (2001b). Seasonal adjustment and signal extraction in economic time series. In *A Course in Time Series Analysis* (Edited by D. Peña, G. C. Tiao and R. S. Tsay), 202-246.

Gómez, V., Maravall, A. and Peña, D. (1999). Missing observations in ARIMA models: Skipping approach versus additive outlier approach. *J. Econom.* **88**, 341-363.

Hannan, E. J. and Rissanen, J. (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika* **69**, 81-94.

Hawkins, S. and Mlodinow, L. (2010). *The Grand Design*. Bantam Books, New York.

Hillmer, S. C. and Tiao, G. C. (1982). An ARIMA-model-based approach to seasonal adjustment. *J. Amer. Statist. Assoc.* **77**, 63-70.

Jenkins, G. M. and Watts, D. G. (1968). *Spectral Analysis and its Applications*. Holden-Day, San Francisco.

Kendall, M. G. and Ord, J. K. (1990). *Time-Series*. Edward Arnold, London.

Lytras, D. P., Feldpausch, R. M. and Bell, W. R. (2007). *Determining Seasonality: A Comparison of Diagnostics from X-12-ARIMA*. U.S. Census Bureau.

Maravall, A., López, R. and Pérez, D. (2015). Reliability of the automatic identification of ARIMA models in program TRAMO. In *Empirical Economic and Financial Research. Theory, Methods and Practice*, Springer International Publishing, Switzerland, 105-122.

Maravall, A. and Pérez, D. (2012). Applying and interpreting model-based seasonal adjustment. The euro-area industrial production series. In *Economic Time Series: Modeling and Seasonality* (Edited by W. R. Bell, S. H. Holan and T. S. McElroy), CRC Press, New York.

Maravall, A. and Pierce, D. A. (1986). The transmission of data noise into policy noise in US monetary control. *Econometrica*, 961-979.

Moore, G. H., Box, G. E. P., Kaitz, H. B., Stephenson, J. A. and Zellner, A. (1981). *Seasonal Adjustment of the Monetary Aggregates: Report of the Committee of Experts on Seasonal Adjustment Techniques*. Washington, D.C.: Board of Governors of the Federal Reserve System.

Pierce, D. A. (1978). Seasonal adjustment when both deterministic and stochastic seasonality are present. In *Seasonal Analysis of Economic Time Series*. (Edited by A. Zellner). Washington, D.C.: U.S. Dept. of Commerce-Bureau of the Census.

Tiao, G. C. and Tsay, R. S. (1983). Consistency properties of least squares estimates of autoregressive parameters in ARMA models. *Ann. Statist.* **10**, 856-871.

Tsay, R. S. (1984). Regression models with time series errors. *J. Amer. Statist. Assoc.* **79**, 118-124.

Tsay, R. S. (1986). Time series model specification in the presence of outliers. *J. Amer. Statist. Assoc.* **81**, 132-141.

United Nations (2011). *Practical Guide to Seasonal Adjustment with Demetra*. United Nations, New York and Geneva.

Bank of Spain, Alcala 48, 28014 Madrid, Spain.

E-mail: amaravall@telefonica.net

Servicios Financieros, Indra Sistemas, Avenida de Bruselas, 35, 28108 Alcobendas, Madrid, Spain.

E-mail: rlopezp@indra.es

Centros de Desarrollo Global, Indra Sistemas, Anabel Segura, 7, 28108 Alcobendas, Madrid, Spain.

E-mail: dperezc@indra.es

(Received May 2014; accepted February 2015)