

Estimating General Proximity Effects using Regression ARIMA

Frank Masci
Version 7.0, 21/4/2006

1. Motivation and Generic Framework

1.1 Goals

This document presents a design to estimate and correct for generic calendar-related proximity effects. For example, Chinese New Year; Easter; Father's Day, Ramadan and any other hitherto unknown effects remaining to be discovered. An outline of previous work and goals were outlined in the proposal document: [1].

In general terms, a "proximity effect" occurs when a "moving holiday" happens to fall close to a period (month or quarter) boundary, thereby causing a systematic difference in the level of activity in the periods that straddle (before and after) the boundary. A "moving holiday" is one whose calendar date changes from year to year. Take Easter for example. When Good Friday falls in early April, retail series usually show excess activity in March due to a build up in sales leading up to the Easter holiday period. At the same time, with the (Australian) four day Easter holiday period falling in April, a decrease in April sales is observed.

The strength of a proximity effect can be estimated by regressing a specific model combined with an ARIMA model against the series data (ie. REGARIMA framework; see [2] for theory). Parameter estimates are then used to correct the series to a level where residual calendar effects become purely regular and periodic in nature. This "seasonality" can then be corrected for using the standard X11 algorithm.

1.2 A Generic Model

We propose the following generic 2-parameter model for the regression mean function:

$$\mu_t = E_b x_b + E_d x_d, \quad (1)$$

where the residuals $z_t = y_t - \mu_t$ can be modelled using ARIMA applied to a (usually) non-stationary input series y_t . x_b and x_d are the "explanatory" variables and defined via a "regression matrix" (see below) and E_b , E_d are the "before or pre-" and the "during or post-" proximity holiday parameters respectively. These parameters are estimated from the REGARIMA fitting procedure.

The explanatory variables are fixed by a regressor model which depend on the moving holiday and also possibly the "theme" (ie. industry) for the series in question. The regression matrix can be constructed from the following generic four parameter model:

$$\begin{aligned}
 x_b &= \left(\frac{n}{w}\right)^{p+1}; & x_d &= \left(\frac{m}{h}\right) \left[\frac{q+1-(m/h)^q}{q} \right] && \text{for "Reference period"} \\
 x_b &= -\left(\frac{n}{w}\right)^{p+1}; & x_d &= -\left(\frac{m}{h}\right) \left[\frac{q+1-(m/h)^q}{q} \right] && \text{for "Reference" period} + 1 \\
 x_b &= 0; & x_d &= 0 && \text{otherwise}
 \end{aligned} \tag{2}$$

See [1] for a derivation. The "reference period" is the specific month or quarter defining the regressor (usually the period before the proximity boundary) and "reference period" + 1 is that proceeding the boundary. The enforced symmetry $x_{b,d}$ ("Ref period") = $-x_{b,d}$ ("Ref period" + 1) ensures that the correction for the proximity effect has no net effect on the final adjusted series over a year.

The four parameters in Equation (2) are defined as:

w = assumed window length in days before holiday date in question;

h = assumed window length in days of holiday period, or, length after holiday date;

p = shape parameter for activity rate over the " w days";

q = shape parameter for activity rate over the " h days". (3)

These have the following limits and ranges. The lower bounds result in physically meaningful explanatory variables, while the "tentative" upper bounds are from existing observations and modelling of the known proximity types in ABS time series:

$$0 < w < \sim 15; \quad w \in \mathbb{N}$$

$$0 < h < \sim 15; \quad h \in \mathbb{N}$$

$$-1 < p < \sim 4; \quad p \in \mathbb{R}$$

$$-1 < q < \sim 4; \quad q \in \mathbb{R}$$

$$\lim_{q \rightarrow 0} x_d = \begin{cases} m/h & \text{for "Reference period"}. \\ -m/h & \text{for "Reference period" + 1.} \end{cases} \tag{4}$$

More precisely, "p" and "q" are power-law indices defining the dependence of the "activity rates" as a function of time before and after the holiday boundary respectively (see schematics in Figure 1). $(p, q) = (1, 1)$ therefore implies a "quadratic-quadratic" model, ie. a quadratic time-dependence in *integrated* activity over the " w " and " h " periods respectively. On the other hand, $(p, q) = (1, 0)$ implies a "quadratic-linear" model. It's important to note that all (monthly or quarterly) periods whose boundaries cross into the windows defined by w or h as a holiday "moves" will be affected by the proximity effect.

The variables " n " and " m " are computed from the date of the specific moving holiday in a given year and are defined as:

n = number of w days falling in "Ref period";

m = number of h days falling in "Ref period". (5)

Figure 1 illustrates the excess activity rate, $d\mu_t / dt$ (ie. excess activity per unit day where μ_t was defined in eqn 1) as a function of time (days) that may be introduced by a proximity effect. The meaning behind the different possible values of the REGARIMA fitted coefficients E_b and E_d (positive, negative or zero) is shown for two flavours of behaviours: $p=0,1$ and $q=0,1$ (ie. linear and quadratic integrated time-dependencies respectively).

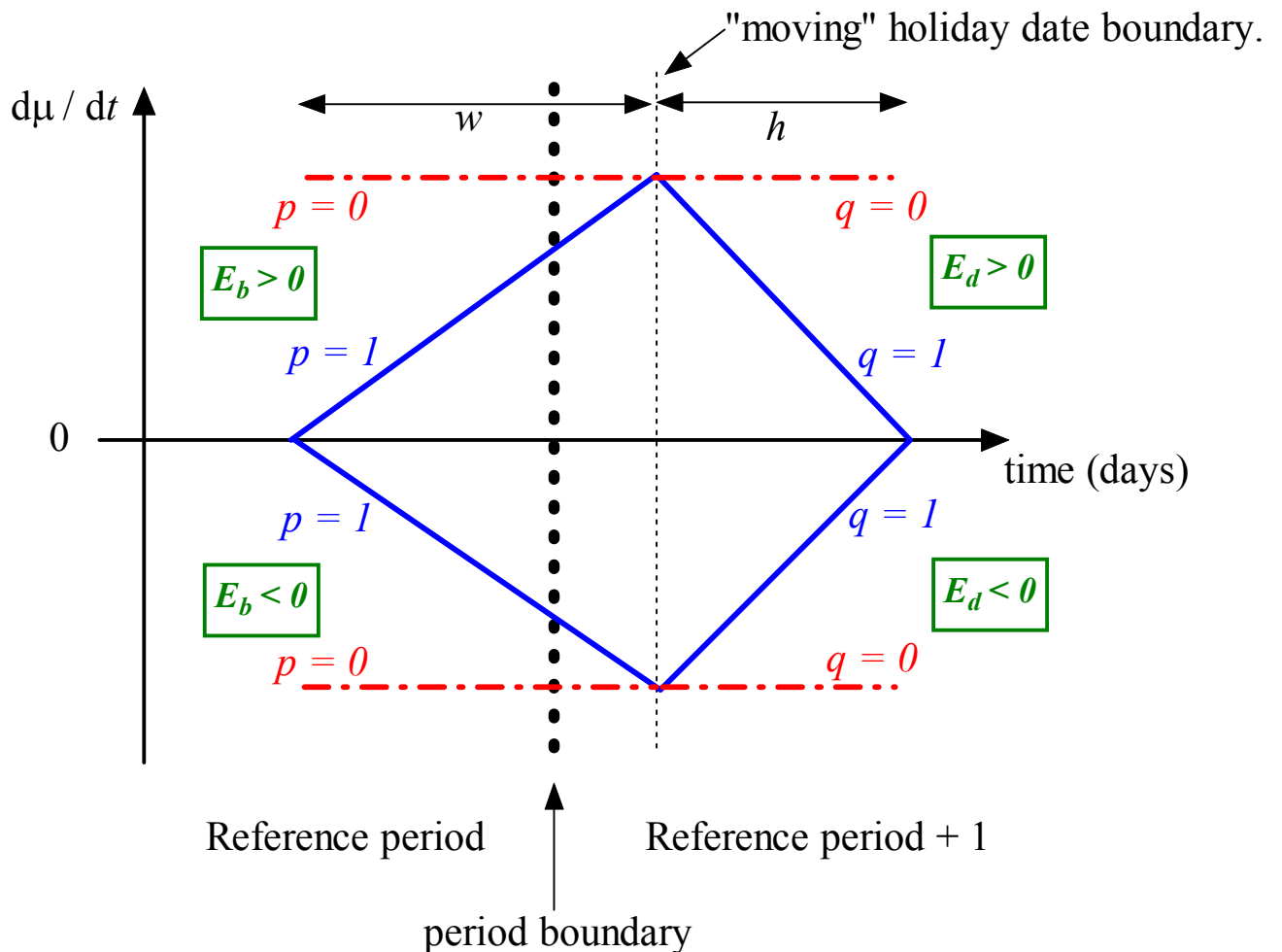


Figure 1: Idealized schematic illustrating possible scenarios that can be introduced by a moving holiday proximity effect. Different scenarios are labelled by various combinations of the sign of non-zero estimates for E_b (before holiday) and E_d (during or after holiday) for $p = 0,1$ and $q = 0,1$. Four model combinations are possible: $(p, q) = (0, 0); (0, 1); (1, 0); (1, 1)$.

1.3 Implications and Interpretations

The parameterisation in equation (2) is appropriate for modelling all four known types of proximity effects: Chinese New Year, Easter, Father's day and Ramadan as applied to most industries examined by the ABS (eg. retail, trade, tourism/travel, hotel accomodation etc). According to this model, we have the following currently observed (hence expected) behaviours for each proximity effect:

Chinese New Year (CNY):

According to a study performed by the TSA section (S:\slides\seminars\MovingHoliday_150705.PRZ), the following parameterisation is found to best represent overseas arrivals and departures data: $p=1$; $q=1$; $4 < w < 8$ ($w \sim 7$); $4 < h < 6$ ($h \sim 6$). This may need to be tuned according to industry type. An estimation of both before (E_b) and after (E_d) effects is required. Specific details on how to compute the CNY proximity regressor are given in section section 5.1.



Easter:

The parameter space to represent the Easter Proximity effect in ABS series was explored in numerous studies. A recent study by F. Masci [📄] found: $0.5 < p < 1.5$; $q=0$; $3 < w < 8$ ($w \sim 7$); $0 < h < 5$ ($h \sim 4$). Note that SEASABS processing assumes the defaults: $p=1$; $q=0$; $w=7$; $h=4$. This effect will require an estimation of both parameters: E_b and E_d . Examples of different behaviours introduced by the Easter proximity are in the retail and tourism or hotel accomodation industries, ie. low and high activity during the Easter holiday period respectively. For retail, we expect $E_b > 0$ and $E_d \leq 0$. For tourism/accomodation we expect for example $E_b > 0$ or $E_b \leq 0$ and $E_d > 0$. Specific details on how to compute the Easter proximity regressor are given in section section 5.2.

Father's Day (FD):

An in depth study of effects from Father's day is yet to be carried out. The current SEASABS model involves a single parameter estimate: E_b (ie. the before Father's day activity is what usually matters) and is represented by $p=1$, and $w=7$. In other words $m = 0$ or $x_d = 0$ (viz. eqn 2) for all periods is implicitly assumed. This makes sense since Father's day itself (the first Sunday in September) and any period thereafter will never fall in August. All post Father's day effects will occur wholly in September and can be corrected using the standard X11 seasonal adjustment procedure. For example, an increase in hardware sales in the week preceeding Father's day (in August) may cause a corresponding decrease (or disinterest) in harware purchases after Father's day (and throughout September). The latter will appear as regular seasonality across a series. Thus, only two parameters are needed to model this effect: p and w . Specific details on how to compute the Father's day proximity regressor are given in section section 5.3.

Ramadan:

According to a study also performed by TSA (S:\slides\seminars\MovingHoliday_150705.PRZ; see also  and ), the following parameterisation is found to best represent overseas arrivals and departures data: $p=1$; $q=1$; $w \sim 6$; $h \sim 6$. This may also need to be tuned according to industry type. An estimation of both before (E_b) and after (E_d) effects is required. Specifically, the date of "Eid Al-Fitr" (on which festivities mark the end of Ramadan) is what defines the "moving holiday boundary" (see Fig. 1). This date can fall in different months (rather than being confined to two months like CNY and Easter). For example, across the years 1993 -> 2005, the date of "Eid Al-Fitr" fell in either Nov, Dec, Jan, Feb or Mar. The generic model defined by Equation (2) can still be used since we can still define an arbitrary "reference" month: either the (pre)ceding month containing a fraction of the "w" window period, or, the (pro)ceeding month containing a fraction of the "h" window period. The only difference with Ramadan is that different pairs of months throughout a series can experience (irregular or non-seasonal) proximity effects. Specific details on how to compute the Ramadan regressor are given in section 5.4.

1.4 Testing for "Balanced" Before and After Proximity Effects

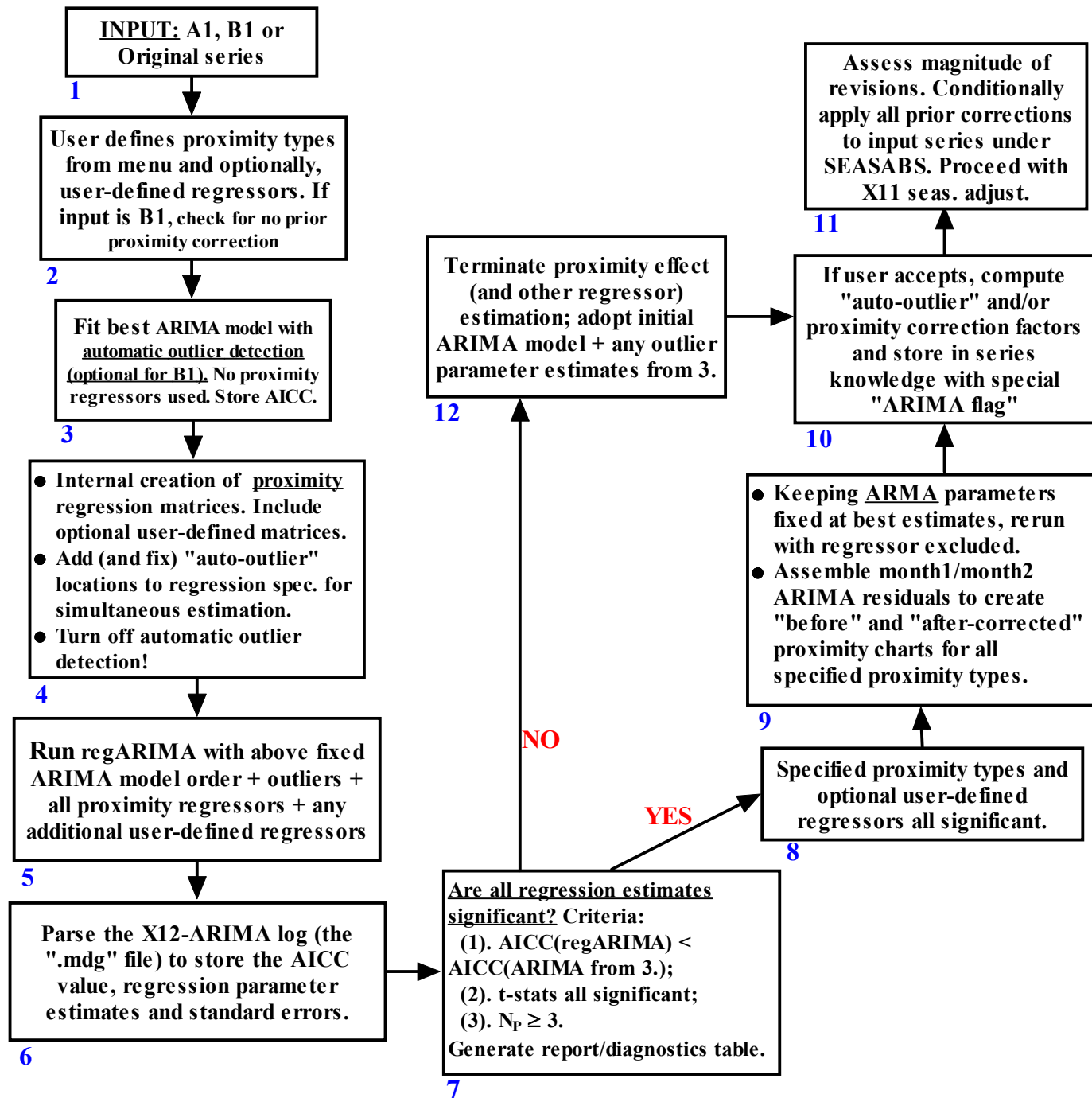
For any proximity effect in general whose regressor can be parameterised in terms of two parameters (eg. E_b and E_d), and regardless of the assumed shapes and window lengths, one may be interested in testing whether the sum " $E_b + E_d$ " differs significantly from zero. In other words, does the before holiday activity balance-out the during/after holiday activity? Put another way, if the before and after periods (windows) fall entirely within a month or quarter, will there be net zero activity (neither an excess or deficit) for that month or quarter? If this is the case, then the moving holiday will not introduce any regular seasonal behaviour for the series in question after its proximity effects are corrected.

2. Algorithmic Flow

The method below unifies the three possible types of input series ($A1$, $B1$ and *Original*) for X12-ARIMA into a generalised framework in correcting for calendar proximity effects. It also allows for optional addition of other user-defined regressors (eg. trading day and/or seasonal regression) for simultaneous estimation. This is a first step to our goal in using model-based approaches to correct for calendar effects. The main features/products from this design are:

- Updates to the current X12-ARIMA interface to select "known" proximity types with parameter specifications for inclusion in regression-ARIMA estimation (eg. see section 3 below).
- Proximity diagnostics summary table (eg. see section 4) containing results for all requested proximity parameter estimates and their significance.
- Proximity charts in terms of ARIMA/regARIMA residuals (before and after correction), eg. see section 6.
- Updates to the series knowledge for all X12-ARIMA specific regression parameter estimates by date and type.

See the **"NOTES"** section that follows for more details on each step (labelled by numbers beside each box).



Notes:

Numbers below refer to steps in the flowchart above.

1. Input series can be either the pre-adjusted SEASABS "A1" (pre-adjusted for ad-hoc user-defined / outlier interventions), prior-corrected SEASABS "B1" or

the unadjusted/uncorrected "original".

2. The user selects the proximity types, their p , q , w , h parameters (if different from defaults), significance critical levels, "split" years and other estimation options from a new pulldown menu on the X12-ARIMA interface called "Proximity Regression" (see proposed design in section 3). This would list the following known types: Chinese new year (CNY); Easter (E), Fathers Day (FD) and Ramadan (R). The user also has the option to import/specify additional (ie. maybe non-proximity related "user-defined") regressors made externally using the existing functionality. If any proximity types were selected and the input series is "B1", then it must be ensured that this series does not contain any of the same prior proximity corrections from SEASABS processing. If so, an error should be issued and further execution halted.

3. The user attempts to fit the best pure ARIMA model with no regressors specified. For the B1 series input case, automatic outlier detection is optional (in addition to prior SEASABS corrections), while for the original input case, it is usually expected. This can be specified by the spec: *outlier{types=all, print=(finaltests)}* which covers AO's, LS's and TC's. There is already a on the X2-ARIMA interface for outlier specification, but it would be good to also include the option *"print=(finaltests)"* in the spec. For selecting the best "pure" ARIMA model, a user can exercise the X12-ARIMA "Automatic modelling" option (that cycles through a list of ARIMA models), with further fine-tuning to ensure no unit roots and other criteria are met. The parameter estimates for this initial ARIMA run, together with its AICC and any auto-outlier locations (ie. their timepoints) should be stored in memory for later use.

4. Proximity regression-matrices are created internally from the user-supplied parameters (from step 2 above) using the formalism in section 5 below. Our goal is to perform a simultaneous estimation of a given set of proximity regressors (and any optional externally imported regressors) with a fixed set of outliers detected in step 3. To do this, we need to copy and fix any automatically detected outlier locations (ie. their timepoints) from the output ".mdg" log file to regression variables in the *"regression{variables(...)}"* spec. E.g., for the different flavours of outliers we may have: *"regression{...variables(ao1972.feb ls2001.nov tc2005.apr)...}"* . Following this, automatic outlier detection now needs to be turned off to avoid distorting regressor estimates in the next step.

5. Regression ARIMA (regARIMA) is now run with the inclusion of:

- (i)** the fixed "best" ARIMA model order from step 3 [ie. fixed (p,d,q)(P,D,Q) specification, not actual parameter estimates];
- (ii)** any fixed "auto-outliers" in the regression{...} spec from step 4;
- (iii)** all internally created proximity regression matrices; and
- (iv)** any additional user-defined, externally created matrices.

Note that if the user specified "split" years for any proximity types in step 2, then new analysis spans must be defined and all parameters estimated (simultaneously) and separately for each span. The log files of interest are those with ".out", ".mdg", ".rcm" and ".rsd" extensions.

6. The X12-ARIMA ".mdg" output log is parsed to pick out the line containing the "AICC" statistic and lines containing all regression parameter names with estimates and standard errors (eg. for outliers, proximity and all other

user-defined regressors). These are all on one line for each regression parameter (see F. Masci for examples). These should be stored in memory for later use.

7. A regression parameter estimate is declared significant (pertaining to either an outlier or any other parameter) if it passes three criteria:

(i) the overall AICC value of the model from which it was estimated is smaller than the initial (no regressor) ARIMA model AICC from step 3 (ie. we want to ensure its inclusion does not degrade the overall model fit);

(ii) the t-statistic for any parameter ($t = P / \sigma$, where P is its estimate and σ its standard error) is significantly different from zero. For proximity parameter estimates, we use the critical levels " x " specified by the user in step 2 (eg. $|t| > \sim 1.96$ for $x = 5\%$ significance and $DOF = N_{\text{data}} - 1 > \sim 200$). For all other parameter estimates (outliers and external user-defined regressors), we assume the default " $x = 5\%$ ";

(iii) if a specific proximity parameter estimate is being assessed, the number of "proximity years", N_p , in the series is ≥ 3 . N_p is regressor model [ie. w, h] dependent. For a given proximity model, it is equal to the number of years across the input series analysis span (or spans if "split" years were specified in step 2) with a corresponding non-zero explanatory variable (x_b or x_d ; see section 5) in the regression-matrix.

- Regardless if these criteria fail or not, a proximity diagnostics/summary table should be generated (see section 4) and made available to the user by clicking on a "peruse proximity results" button.
- If either criteria (i), (ii) or (iii) fail (for any "split-defined" analysis spans), and if it's the first time regARIMA proximity correction was run (by checking the series knowledge), a warning should be issued summarising the problem encountered with a "clickable" option to either adopt the initial (step 2) ARIMA + any auto-outlier estimates and move to step 12, or, to terminate and re-run regARIMA with the insignificant parameters excluded.
- If a previous regARIMA proximity correction exists in the series knowledge (ie. the last run at time t) and has $|t\text{-stat}(i)_{\text{lastrun}}| \geq t_{\text{crit}}(x\%)$, but, for the same parameter i in the new run $|t\text{-stat}(i)_{\text{newrun}}| < t_{\text{crit}}(x\%)$, the proximity effect may cease to be significant. In this case, the user should be warned that a "split year" separating the series at times $\leq t$ from times $t+1..$ may need to be declared (in step 2). New parameters must then be re-estimated for all new "split" analysis spans.
- If all the above three criteria pass (for all parameters), we declare that all regression estimates are significant and move to step 9.

8. This is just a statement if the criteria from step 7 are satisfied. Nothing to be explicitly done here.

9. Given the regARIMA model (that passed the criteria in step 7), the X12-ARIMA program is now re-run by keeping the ARMA parameters fixed at their best estimates and with no regressors specified. This is for the purpose of creating the "before-correction proximity chart(s)". The "after-corrected proximity chart(s)" are created using the best regression model residuals from the "best" model in step 7 (see examples in section 6 below). These charts will be made exclusively from (reg)ARIMA model residuals on the fly when requested by the user (ie. a new button on the X12-ARIMA output results interface). The monthly model residuals (eg. "Jan/Feb" for Chinese New Year, "Mar/Apr" for

Easter, "Aug/Sep" for Father's Day and arbitrary "month_bef/month_aft" for Ramadan) are assembled from the output ".rsd" files. These files are produced by including the "save=(residuals) " option in the "estimate{...} " spec. To fix a set of ARMA parameters at their best estimates (for creating the "before-correction prox. chart"), they need to be copied from the ".mdg" summary file to the "arima{...} " spec and written in the format "ar=(ar(1)f, ar(2)f, etc...); ma=(ma(1)f, ma(2)f, etc...) ". The "f " in these lists is a character suffix that needs to be appended to each value and means that the parameter is to be held fixed during estimation. Having done this, all regressor specifications are now disabled and X12-ARIMA is rerun. It's important to note that these new proximity charts will only be used by the TSA section for diagnosis and revision assessment. The existing SEASABS charts made using the D13 irregulars must remain.

10. The regARIMA parameter estimates are converted to correction factors (as already done if a user selects "Accept factors" on the X12-ARIMA output products interface) and stored in the series knowledge. We also want to store the actual proximity parameter estimates and their corresponding t-values (to assess revisions in the next step). We will want to discriminate between specific ARIMA corrections and standard SEASABS corrections in the series knowledge (maybe using a flag in a new column) and visible on the "Prior Factors..." SEASABS interface. Users should be able to enable/disable any of these in future processing. We will want to separate out each type of regARIMA detected outlier and each proximity parameter estimate with dates for the former and applicable time-spans for the latter.

11. Before applying new proximity regARIMA corrections to the input series (after combining these with all other prior factors to populate the "Special Prior Factors..." table in SEASABS), we want to ensure that it makes sense to do so. We take a conservative approach and only perform revisions to stored proximity estimates and hence seasonally adjusted series **if either:**

(i) it's the first time regARIMA proximity corrections were computed and all criteria in step 7 were satisfied; **or**

(ii) if there were any significant historical revisions to the original series and all criteria in step 7 were satisfied; **or**

(iii) if timepoint t+1 is defined as a "proximity year" according to the specific regression model and moving holiday in question, and of course if all criteria in step 7 were satisfied.

Note that revisions due to a regression parameter becoming insignificant all of a sudden was covered in step 7. In that step, a recommendation is made to the user to define a "split" analysis span for subsequent regARIMA proximity estimates.

12. Termination for diagnosis of a proximity effect occurs if any of the criteria in step 7 were not satisfied. See explanation in step 7 above for processing options from here.

3. Updates to X12-ARIMA Interface

Run X-12-ARIMA

Series: LIQUORWITHSIMOUTLIERSUC94SI903039

Input Series: Original Pre-adjusted (A1) Prior adjusted (B1)

Model span: Whole series span No more than the last 15 years Specified span: September 1990 To: August 2005

Automatic modelling

ARIMA Model: (2 1 0)(1 1 1)

Current ARIMA model: (2 1 0)(1 1 1)
 Parameters:
 Non seasonal AR lag 1 -0.707719
 Non seasonal AR lag 2 -0.396340
 Seasonal AR lag 12 0.284123

Regression Matrix: Regressor_wp_3.3 Show: Regression Matrices

Current General Regression corrections:
 xb_whp_703 0.0209954
 ao1 -0.103817
 ao2 -0.214920
 ao3 -0.120493

Additional X-12-ARIMA Specifications: Identify Outlier detection

Place this here

Proximity Regression

Chinese New Year Easter Father's Day Ramadan

e.g.

Proximity Application

Estimate and apply if significant Estimate but do not apply

"Split" spans for regression

Whole series (analysis) span Specified "splits" (prd/yr boundaries):
 Jan/1990; Nov/2000

Parameters

Pre-holiday window (*w*): 7 day(s)
 Post/during-holiday window (*h*): 6 day(s)
 Power-law index *p* over "*w*" days: 1
 Power-law index *q* over "*h*" days: 1
 Sig- level for $H_0(E_b=0 \text{ or } E_d=0)$: 5 %
 Sig- level for $H_0(E_b+E_d=0)$: 5 %

Default for this field (initialized on start-up) is "Estimate but do not apply".

Semi-colon separated list of period/year boundaries separating split analysis spans. In this example, proximity regression is to be performed in three separate spans:

- (i) start-span→Jan/1990;
- (ii) Feb/1990→Nov/2000;
- (iii) Dec/2000→end-span

Default (initialized on start-up) is: "Whole series (analysis span)".

Note that all these should be initialized to default values for each proximity effect (see suggested starting defaults in section 1.3). For Father's Day proximity, only one regressor applies so that only the "*w*", "*p*" and "Sig-level $H_0(E_b=0)$ " fields should appear on this panel.

Parameters

Pre-holiday window (w):	7	day(s)
Post/during-holiday window (h):	6	day(s)
Power-law index p over " w " days:	1	
Power-law index q over " h " days:	1	
Sig- level for $H_0(E_b=0$ or $E_d=0)$:	5	%
Sig- level for $H_0(E_b+E_d=0)$:	5	%

Default (initialized on start-up) is: "Whole series (analysis span)".

Note that all these should be initialized to default values for each proximity effect (see suggested starting defaults in section 1.3). For Father's Day proximity, only one regressor applies so that only the " w ", " p " and "Sig-level $H_0(E_b=0)$ " fields should appear on this panel.

Figure 2: Additions to existing X12-ARIMA interface.

4. Proximity Results/Summary Table (step 7)

- The table below is an example summary output showing the test results of all proximity types specified on input (ie. checked by the user on the interface in Fig. 2). Also, if "split" analysis spans were specified, the results for each span (in separate rows) should be shown. The example below shows two analysis spans for every proximity type, although according to the above design, there is no limit to the number of spans the user can specify.
- Note that only one parameter " E_b " is relevant for Father's day. For this case, all columns (with missing values) other than " N_p ", " E_b ", " $\sigma(E_b)$ ", "AICC", " Δ AICC", " $t(E_b)$ " and "P-value $H_0(E_b=0)$ " should specify the string "N/A".
- We envisage this table to be made available to the user by clicking on a "peruse proximity results" button on the X12-ARIMA "output products" interface.

Proximity Effect 	N_p	E_b	E_d	$\sigma(E_b)$	$\sigma(E_d)$	AICC	Δ AICC	$t(E_b)$	$t(E_d)$	$t(E_b+E_d)$	P-value $H_0(E_b=0)$	P-value $H_0(E_d=0)$	P-value $H_0(E_b+E_d=0)$
CNY <span1>: <span2>:													
Easter <span1>: <span2>:													
Father's D. <span1>: <span2>:													
Ramadan <span1>: <span2>:													

Column descriptions:

- First column lists the proximity type and all user-specified spans thereunder (default is full analysis span for each type).
- Second column is number of "proximity years" in relevant span (see step 7 above for definition).
- E_b and E_d are the generic "before" and "during" holiday regression parameter estimates from the ".mdg" output log.
- $\sigma(E_b)$ and $\sigma(E_d)$ are corresponding standard errors also from the ".mdg" output log.

- The **AICC** statistic is next and also from the ".mdg" output log.
- ΔAICC = difference between the initial (no regressors) AICC statistic (from step 2) and the final model AICC, i.e. $\Delta\text{AICC} = \text{AICC}(\text{initial}) - \text{AICC}(\text{regressor model})$.
- The t-statistics: $\mathbf{t}(E_b)$ and $\mathbf{t}(E_d)$ are computed from the ratios $E_b/\sigma(E_b)$; $E_d/\sigma(E_d)$ respectively.
- $\mathbf{t}(E_b + E_d)$ represents the t-statistic for the sum of the "before" and "after" Easter holiday parameter estimates, computed from:

$$t(E_b + E_d) = \frac{E_b + E_d}{\sqrt{\sigma_{Eb}^2 + \sigma_{Ed}^2 + 2\text{cov}(E_b, E_d)}},$$

where the variances are given by the squares of the standard errors above and the covariance term can be obtained from the output ".rcm" file which contains the full error-covariance matrix of all regression parameters. The ".rcm" file is produced by specifying "regcmatrix" in the "estimate{...}" spec prior to the X12-ARIMA run. See section 1.4 for interpretation.

- The last three columns are probability measures (P-values) for three scenarios derived from the t-stat values in the preceding three columns. These test for three different holiday effects: **(i)** "no before holiday effect" (parameter estimate $E_b=0$); **(ii)** "no after or during holiday effects" (parameter estimate $E_d=0$) and **(iii)** "whether the before and after (or during) holiday effects negate (or cancel) each other" ($E_b + E_d = 0$) - see section 1.4 for interpretation.

5. Internal Computation of Proximity Regression Matrices (step 4)

Below we describe the method to compute the "explanatory variable regression matrix" for each proximity type. This is based in the formalism presented in section 1.2 (specifically Equation 2). It is assumed values for the model parameters: w , h , p and q have been specified by the user. The regression matrix [eqn. 2] can then be computed once the variables " n " and " m " are known. This requires knowledge of the actual date of the moving holiday for all relevant years. References to moving holiday dates are given below. It is suggested these dates be stored as look-up tables in SEASABS (as currently done for the Easter dates).

5.1 Chinese New Year (CNY)

The start date of CNY was derived using a piece of mathematica code to generate a table of dates for years 1950 - 2050: "

S:\regressors\Chinese_New_Year!\CNY_1950_2050.dat". The first several lines of this table are:

<u>yr</u>	<u>mnth</u>	<u>day</u>
1950	2	17
1951	2	6
1952	1	27
1953	2	14
1954	2	3
1955	1	24

1956	2	12
1957	1	31
1958	2	18
.		
.		

First, the "pre-holiday" window w is defined as the number of days up to the last day before (Chinese) New Year's day and second, the "post/during-holiday" window h is defined as the number of days starting at (ie. including) New Year's day onwards. Computation of the variables n ("number of w days" that fall in the January reference month) and m ("number of h days" that fall in January) [eg. see Fig. 1], entails counting the number of whole days that fall in January from each of these windows respectively. The regression matrix can then be computed for a given series span (and all periods within) using Equation 2 and stored internally. An example regression matrix using $p=q=1$; $w=7$; $h=6$ and the above table is given in: "[S:\regressors\Chinese_New_Year!\QQ_76.txt](#)".

5.2 Easter

The date of Easter Sunday is what's relevant here. These dates are available in a table for years 1900 - 2100: "[S:\regressors\dates\EasterDatesInp.txt](#)". It is unknown at the time of writing what range SEASABS currently stores. The first several lines of this table are:

<u>yr</u>	<u>mnth</u>	<u>day</u>
1900	4	15
1901	4	7
1902	3	30
1903	4	12
1904	4	3
1905	4	23
1906	4	15
1907	3	31
1908	4	19
.		
.		

First, the "pre-holiday" window w is defined as the number of days up to the last day before Good Friday (ie. the Thursday) and second, the "post/during-holiday" window h is defined as the number of days starting at (ie. including) Good Friday onwards (typically $h = 4$). Computation of the variables n ("number of w days" that fall in the March reference month) and m ("number of h days" that fall in March) [eg. see Fig. 1], entails counting the number of whole days that fall in March from each of these windows respectively. The regression matrix can then be computed for a given series span (and all periods within) using Equation 2 and stored internally. An example regression matrix using $p=1$; $q=0$; $w=7$; $h=4$ and the above table is given in:

"[S:\regressors\EasterProximity\EasterRegressor_whp_742.txt](#)".

5.3 Father's Day

The date of the first Sunday in September is what's relevant here. These dates are available in a table for years 1900 - 2100: "[S:\regressors\dates\FathersDayDatesInp.txt](#)". It is unknown at the time of writing what range SEASABS currently stores. The first several lines of this table are:

yr	mnt	day
1900	9	2
1901	9	1
1902	9	7
1903	9	6
1904	9	4
1905	9	3
1906	9	2
1907	9	1
1908	9	6
.		
.		

The "pre-holiday" window w is defined as the number of days up to the last day before Father's day (ie. the Saturday), and as described in section 1.3, there is no explicit "post/during-holiday" window h in this model. Thus, the Father's day proximity only requires computation of the variable n ("number of w days" that fall in the August reference month) to yield the " x_b " explanatory variable in Equation 2. Computation of n entails counting the number of whole days that fall in August from the w window. The regression matrix can then be computed for a given series span (and all periods within) using Equation 2 and stored internally. An example regression matrix using $p=1$; $w=7$; and the above table is given in: "[S:\regressors\Father'sDay\Regressor_wp_7.2.txt](#)".

5.4 Ramadan

The timing of this holiday is dependent on the Islamic calendar and can occur in any period during the year which is unlike CNY and Easter which will only ever fall in Jan/ Feb and Mar/Apr respectively. Due to this fact, the regressors are constructed differently. The date of "Eid Al-Fitr" (marking the end of Ramadan and the start of festivities) is what matters. These dates are available in a table for years 1932 - 2173: "[S:\regressors\Ramadan\Ram_dates_1932_2173.dat](#)". The first several lines of this table are:

yr	mnt	day
1932	2	7
1933	1	26
1934	1	15
1935	1	5
1935	12	25
1936	12	13
1937	12	3
1938	11	22
.		
.		

As above, the "pre-holiday" window w is defined as the number of days up to the last day before "Eid Al-Fitr", and the "post/during-holiday" window h is defined as the number of days starting at (ie. including) "Eid Al-Fitr" onwards. There are three possible cases:

1. If the date of "Eid Al-Fitr" falls close to the beginning of say $month_i$, [$i = 1..12$], ie. such that the month start boundary overlaps with the " w " window period, then computation of the variable n entails counting the number of whole " w " window days that fall in the preceding month, ie. $month_{i-1}$. From this, the

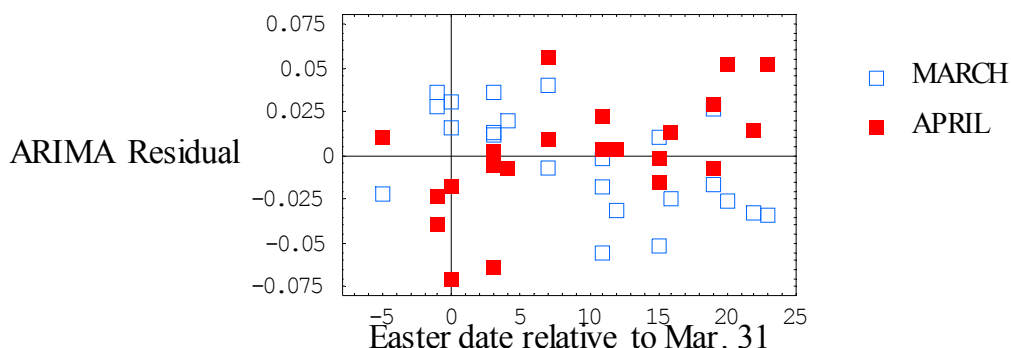
explanatory variables " $x_b(month_i)$ " and " $x_b(month_{i-1}) = -x_b(month_i)$ " can be computed using Equation 2. Under this scenario, the variable m (=number of " h " days that fall in $month_{i-1}$), and hence " $x_d(month_{i-1})$ ", " $x_d(month_i)$ " are precisely zero.

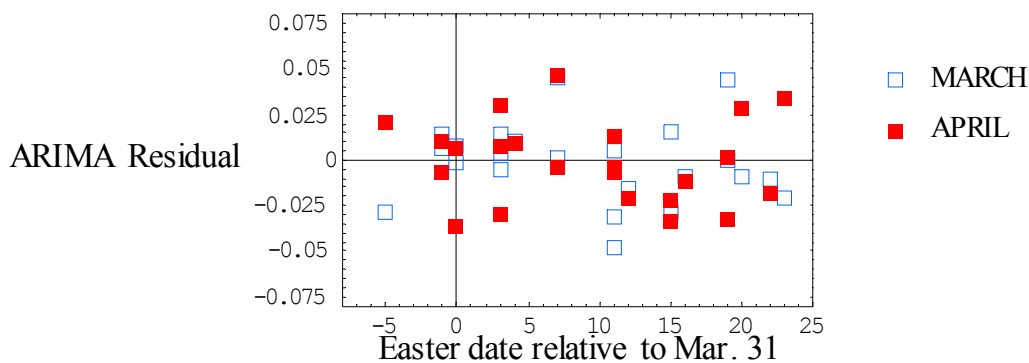
2. If the date of "Eid Al-Fitr" falls close to the end of say $month_i$ [$i = 1..12$], ie. such that the month end boundary overlaps with the " h " window period, then computation of the variable m entails counting the number of whole " h " window days that fall in the proceeding month, ie. $month_{i+1}$. From this, the explanatory variables " $x_d(month_i)$ " and " $x_d(month_{i+1}) = -x_d(month_i)$ " can be computed using Equation 2. Under this scenario, the variable n (=number of " w " days that fall in $month_{i+1}$), and hence " $x_b(month_{i+1})$ ", " $x_b(month_i)$ " are precisely zero.
3. If the date of "Eid Al-Fitr" falls in the middle of say $month_i$ [$i = 1..12$], ie. such that the " w " and " h " windows fall entirely within that month, then both explanatory variables take on values of one. This is to indicate that the total proportion of increased and/or decreased activity falls in that one month. In other words, under this scenario: " $x_b(month_i) = 1$ "; " $x_b(month_{i-1}) = -1$ "; " $x_d(month_i) = 1$ "; " $x_d(month_{i+1}) = -1$ " and $[x_b, x_d] = [0, 0]$ for all other months in that year.

An example regression matrix using $p=q=1$; $w=h=6$ and the above table is given in: "S:\regressors\Ramadan\Ramadan_66.txt".

6. Generalised "Before" and "After" Proximity charts (step 9)

The example charts below pertain to the Easter proximity and are purely illustrative. These were generated by first estimating the "best" ARIMA + proximity regressor model giving the bottom chart. The Easter proximity regression component was then turned off with the ARIMA model parameters fixed at their "best" initial estimates to generate the top chart.





6.1 Notes on the representation of known proximity effects

- For Chinese New Year, the horizontal-axis tick marks should show the date of Chinese New Year's Day. The Jan/Feb (reg)ARIMA residual values should be labelled with the year.
- For Easter, the horizontal-axis tick marks should show the date of Easter Sunday, just like the standard SEASABS charts. The Mar/Apr (reg)ARIMA residual values should be labelled with the year.
- For Father's day, the horizontal-axis tick marks should show the date of Father's day date (first Sunday in September), just like the standard SEASABS charts. The Aug/Sep (reg)ARIMA residual values should be labelled with the year.
- For Ramadan, things are a little more complicated since there are no fixed pairs of months within which the date of "Eid Al-Fitr" moves. For this effect, we would like to represent the proximity chart in terms of an arbitrary "*month_1*" and "*month_2*", where $month_1 = month_i$ (from section 5.4), ie. that month which contains the "Eid Al-Fitr" date, and $month_2 =$ either $month_{i-1}$ or $month_{i+1}$, ie. that month which contains a non-zero proportion of (excess or deficit) activity due to spillage of either the w or h windows into that month respectively. More precisely, $month_2 = month_{i-1}$ if "Eid Al-Fitr" falls close to the beginning of $month_i$, [$j = 1..12$], (ie. such that the month start boundary overlaps with the " w " window period), or, $month_2 = month_{i+1}$ if "Eid Al-Fitr" falls close to the end of $month_i$, (ie. such that the month end boundary overlaps with the " h " window period). If both the w and h windows fall in the same month, then we define $month_2 = month_{i-1}$. The horizontal-axis tick marks should be in terms of the number of \pm days the "Eid Al-Fitr" date falls relative to the closest month boundary. If "Eid Al-Fitr" falls late in a month, then this measure is negative, otherwise, it is positive. If it falls on the last/start day of a month, then this measure is -1/+1. If "Eid Al-Fitr" splits a month exactly, then the positive measure should be taken. Each (reg)ARIMA residual value (corresponding to "*month_1*" and "*month_2*") should be labelled with the year.

