

November 2004

Preliminary

**DATA QUALITY CONTROL DATA
THROUGH OPTIMAL FILTERING**

Agustín Maravall

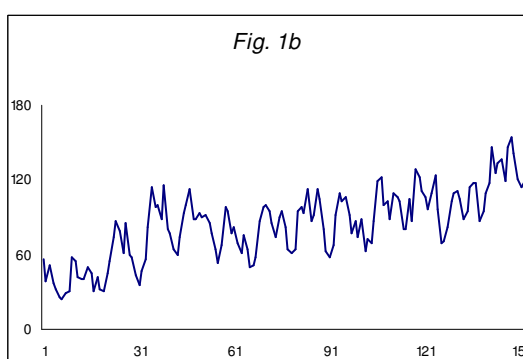
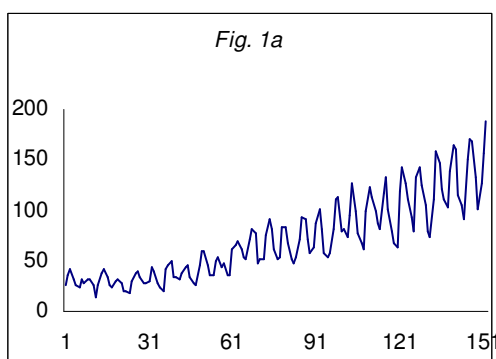
Bank of Spain

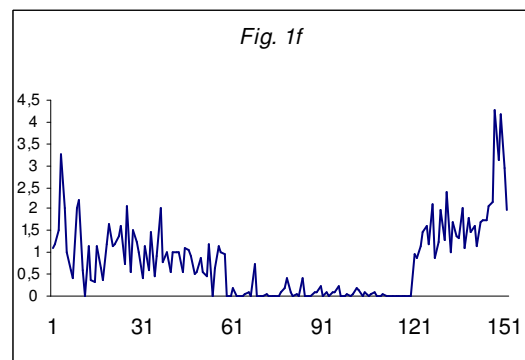
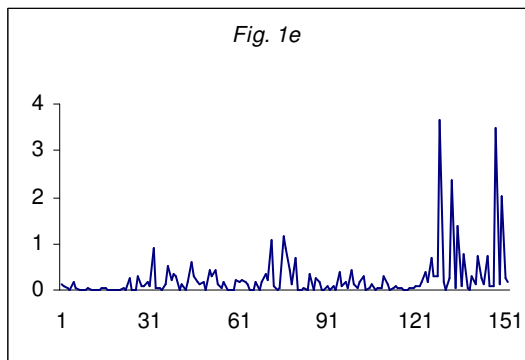
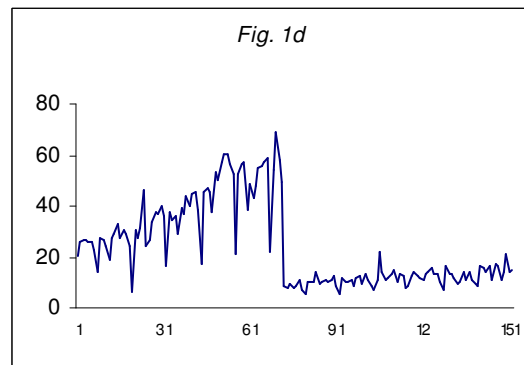
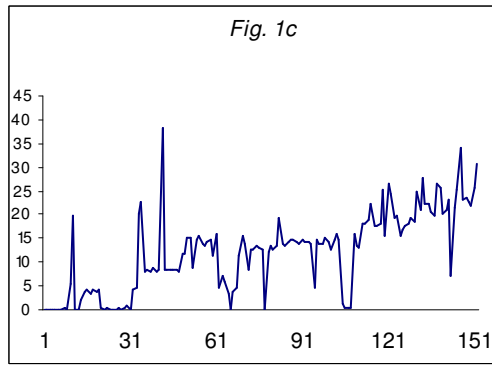
An application of a program for error detection in incoming data for data bases of time series is presented. It provides a sensible answer to the question: when new data is reported, which of the new observations are likely to contain an error? The observation is suspicious if it is far from what could have been expected from the past history of the variable.

Without considering the new observation, and using a fully automatic procedure, a REGARIMA model is identified for the series, which may have missing observations and be contaminated by outliers. Calendar and, more generally, regression effects can also be present. The new observation is compared to the (out-of-sample) forecast obtained with the model and, if the forecast error is clearly unreasonable, the series is considered suspicious (the automatic outlier detection facility can be, of course, exploited to investigate possible former errors).

The program, named TERROR, is a particular application of program TRAMO, freely available (for different platforms) at the Bank of Spain web site (www.bde.es), together with detailed documentation. The program is reliable and efficient for very large scale applications. As an example of its performance, we present a summary of the results for a set of 376 monthly series of 152 observations each. The series are the Spanish Customs registers of exports and imports, classified by type of product and use. TERROR is applied to detect possible errors in the last observation (August 2004), with all parameters taking their default value, and adding pre-tests for Trading Day and Easter effects.

The series in the set are highly volatile. A few have just some non-zero values and they will be set aside by the program. A few display a somewhat regular structure (Figure 1a); the vast majority display a non-regular behaviour that can still be modelled (Figure 1b), and often show the need for outlier adjustment (Figure 1c) or present obvious regime shifts (Figure 1d). On occasion, the series has a clear non-linear structure (Figure 1e), or evidences the forecasting difficulty (Figure 1f).





Some of the output produced by the program is illustrated. First, the list of series that are suspicious of containing an error. This list can be increased or decreased by changing the values of k_1 and k_2 (the threshold for the t-values associated with the standardized forecast error). When $4 < |t| < 5$, the series is classified as containing a “possible” error in the last reported observation; if $|t| > 5$, the series is classified as “likely” to contain an error.

TABLE 1: TERROR TSW SERIES LIST

Date tested: 8/04

Input Parameters: itrads=-2 ieast=-1 terror= 1 k1= 4.000 k2= 5.000

SERIES TITLE	New Value	Forecast	Log (New Value)	Log (Forecast)	Diff.	StdDev	T-Value	Results
39M15AI	1.562.150	1.030.925	2.748.648	2.328.053	0.4205949	0.0998827	4.21	Possible
148M342C	0.8835000	1.800.792	-	-	-0.917292	0.2248348	-4.08	Possible
153M35AI	0.1854000	1.372.065	-	-	-1.186.665	0.2086855	-5.69	Likely
162M35DE	1.907.600	0.6169537	0.6458459	-0.518068	1.163.914	0.2649798	4.39	Possible
167M36AI	1.453.360	2.990.993	-	-	-1.537.633	2.400.159	-6.41	Likely
186M98I	8.721.800	3.038.279	-	-	-2.166.099	2.742.708	-7.90	Likely
208X11BT	2.594.850	2.987.277	-	-	-3.924.270	0.5645003	-6.95	Likely
277X252C	3.450.320	2.540.946	-	-	9.093.742	1.962.164	4.63	Possible
309X30C	0.1400000	0.0037923	-	-	0.1362077	0.0072173	18.87	Likely
318X32AE	4.745.610	6.925.698	-	-	-2.180.088	3.149.316	-6.92	Likely

Series that did not match TERROR memory constraints (Not enough observations or Too Many M.O.):

SERIES TITLE: 21M12I, 81M24BI, 209X12I, 260X22AE, 263X23C, 321X32BC

Summary Statistics

376 Series were tested.

4 Releases were suspicious (possibly wrong).

6 Releases were very suspicious (likely wrong).

0 Series produced a Run-Time EXCEPTION.

6 Series did not match TERROR memory constraints.

360 Series passed the plausibility tests.

The aggregate results of the automatic modelling for the in-sample period are summarized. The following table presents some of them. (For a description, see Caporello and Maravall, 2003.)

TABLE 2: AGGREGATE RESULTS

Input parameters: Terror = 1, ITRAD = -2, IEAST = -1, K1 = 4 (D), K2 = 5 (D).

Series in file/processed: 376/370 . (For 6 series: not enough non-zero values.)

Frequency of observation: monthly; NZ = 151.

- **LEVELS/LOGS:** 134/236

- **DIFFERENCING:**

NONE	ONLY REGULAR	ONLY SEASONAL	BOTH REGULAR AND SEASONAL
11	60	33	266

- **AVERAGE NUMBER OF ARIMA PARAMETERS PER SERIES:** 2.1

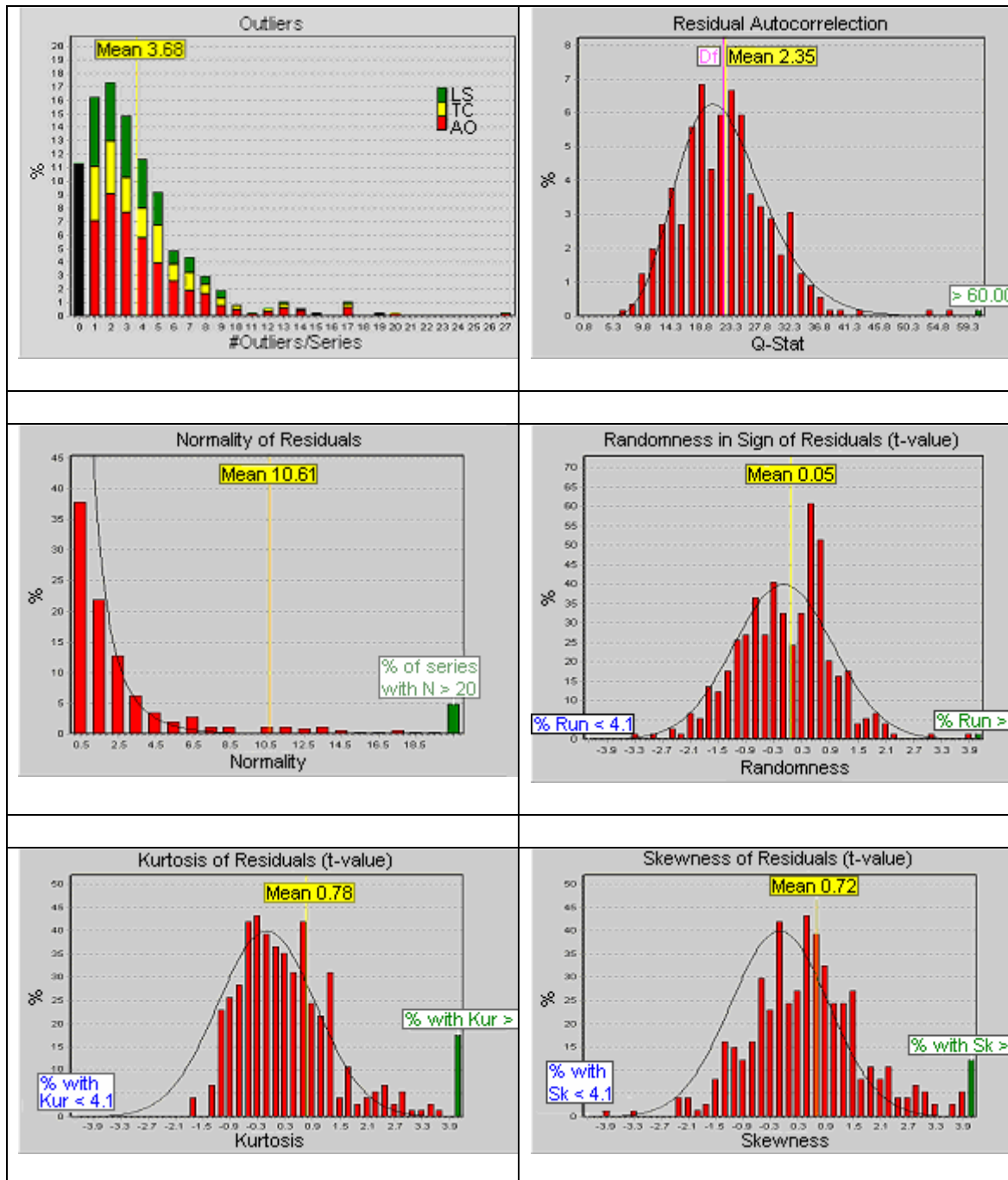
- **“MODE” MODEL:** (0 1 1) (0 1 1)₁₂

- **AVERAGE NUMBER OF OUTLIERS PER SERIES:** TOTAL [AO TC LS]
3.7 [1.9 0.9 0.9]

- **PERCENT OF SERIES WITH CALENDAR EFFECT:** TOTAL [TD EE]
64.9 [61.6 24.1]

DIAGNOSTICS:	MEAN	APPROX. 1% CV	% OF SERIES THAT PASS THE TEST (99%)
Residual AC (χ^2_{22})	23.3	40.30	99.5
Normality (χ^2_2)	5.7	9.21	92.7
Skewness (t)	0.6	2.58	93.2
Kurtosis (t)	0.4	2.58	96.5
Residual Seasonality (χ^2_2)	3.2	9.21	98.9
Nonlinearity (χ^2_{24})	26.6	43.00	91.4
Random Residuals sign (t)	0.0	2.58	98.1

Several graphs with histograms are available. Some are displayed in Figure 2.



Individual results for each series are provided: First, the main statistics describing the fit; second, a summary of the deterministic effects; third, the estimates of the ARIMA parameters, the outlier effects, and the calendar effects. The first rows of these tables are shown below.

TABLE 3: INDIVIDUAL RESULTS.

1. MODEL AND DIAGNOSTICS

n	TITLE	Lam	Mean	P	D	Q	BP	BD	BQ	SE(res)	BIC	Q-val	N-test	SK(t)	KUR(t)	QS	Q2	RUNS
1	"1M01AC"	1	0	0	1	1	0	1	1	10.03394	4.69708	16.93	2.18	0.907	-1.17	0.	29.79	-0.51
2	"2M01AI"	0	0	0	1	1	0	1	1	0.1420778	-3.73299	18.04	2.12	0.610	1.32	0.034	21.74	-0.35
3	"3M01AT"	0	0	0	1	1	0	1	1	0.1180897	-4.13100	17.34	2.84	1.03	1.33	0.042	9.792	1.56
4	"4M01BC"	0	1	1	0	0	0	0	0	0.2042107	-3.04474	23.20	4.63	-1.36	1.67	2.52	43.60	-0.33
5	"5M01BE"	1	0	1	1	1	0	1	1	0.8975408	-0.05725	27.78	0.007	-0.01	0.080	0.280	14.73	0.689

Notes: Lam = 1 levels, 0 logs; (PDQ) (BP BD BQ) = Orders of ARIMA model; Q-val: Residual autocorrelation; N: Normality; SK: Skewness; KUR: Kurtosis; QS: Residual seasonality; Q2: Nonlinearity; RUNS: randomness in signs.

2. ARMA PARAMETER

n	TITLE	PHI1	(t)	"..."	TH1	(t)	"..."	BTH	(t)
1	"1M01AC"	-	(-)		-0.48807	(-6.4)		-0.74292	(8.7)
2	"2M01AI"	-	(-)		-0.55551	(-7.5)		-0.62906	(7.4)
3	"3M01AT"	-	(-)		-0.44024	(-5.7)		-0.66687	(8.2)
4	"4M01BC"	-0.50337	(7.2)		-	(-)		-	(-)
5	"5M01BE"	-0.49282	(11.)		-0.78988	(-18.)		-0.72391	(8.3)

3. DETERMINISTIC EFFECTS

n	TITLE	TD	EE	#OUT	AO	TC	LS
1	"1M01AC"	1	0	0	0	0	0
2	"2M01AI"	1	0	3	1	1	1
3	"3M01AT"	1	0	2	1	1	0
4	"4M01BC"	0	0	3	1	1	1
5	"5M01BE"	0	0	2	1	0	1

4. OUTLIERS

1	"1M01AC"	-----	-----	-----
2	"2M01AI"	AO01(0800, -4.06)	TC01(1292, -4.11)	LS01(0397, 3.47)
3	"3M01AT"	AO01(0800, -4.02)	TC01(1292, -4.11)	
4	"4M01BC"	AO01(0192, -4.20)	TC01(0393, 4.09)	LS01(0196, 30.69)
5	"5M01BE"	AO01(0200, 8.06)	LS01(1200, -3.98)	

Notes: (MMYY, t) = (Month Year, t-value)

5. CALENDAR EFFECT

n	TITLE	TD1	(t)	LY	(t)	EE	(t)
1	"1M01AC"	1.00385	(4.8)	-	(-)	-	(-)

)
2	"2M01AI"	0.012837	(4.2)	- (-)	- (-)
3	"3M01AT"	0.012424	(5.2)	- (-)	- (-)
4	"4M01BC"	-	(-)	- (-)	- (-)
5	"5M01BE"	-	(-)	- (-)	- (-)

Execution of the program on the full set of 376 series with 152 observations took less than 25" in a laptop with a 1.9 Gh processor and a RAM of 512. The conclusion is that, in one day many thousands of series can be treated with a simple PC in a reliable manner. The program can be a useful additional tool in quality control of data.

REFERENCES (available at www.bde.es)

CAPORELLO, G. and MARAVALL, A. (2003), "A Tool for Quality Control of Time Series Data. Program TERROR", Documento Ocasional 0301, Servicio de Estudios, Banco de España.