Working Paper.
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28-30 April 2014)

Topic (i): Selective editing / macro editing

Selective Editing Techniques and Seasonal Adjustment.

Prepared by Thomas Balcone, Antonia Bertin, Marie Cordier-Villoing, Dominique Ladiray,
INSEE, France

I. Introduction

1. In 2009, the French National Institute of Statistics and Economic Studies launched a large redesign of its short-term statistic (STS) production system. Many economic indicators were concerned: Business tendency surveys, Industrial production indexes, Turnover indexes, Production price indexes, etc. One of the main objectives of this project is the definition and the sharing of approved methodologies. This work was mainly done in the GSBPM framework and, as far as statistical methodology is concerned, following the Edimbus manual recommendations (Guggemos, 2010; Luzi et al, 2005).

2. Section 2 quickly presents the global and common organisation of the edit and imputation techniques in the project. Section 3 puts the emphasis on the application of selective editing techniques in the seasonal adjustment process. During this important step, which occurs just a few days before the publication of the final results, the producer has very few time to analyze the series and eventually to react to and correct for unexpected problems. Selective editing techniques are used to help him, based on both the quality of the adjustment and the weight of the series in the aggregates.

II. Data Editing and Imputation of Raw Data

3. In short term statistics, the main outputs are time series (indexes), and growth rates. Any evaluation of the quality is therefore based on both revisions and likelihood of the yearly, quarterly or monthly evolutions. Revisions are necessary and can be the consequence of late responses but high revisions are often suspicious, as are high yearly growth rates. Apart this very specificity, data editing in the STS production system is based on a usual two-step process combining micro-editing and selective-editing.

4. In a first step, raw data are automatically controlled at the micro-level, using a set of classical micro-editing techniques. Data with a very bad quality as well as non-responses are automatically corrected by imputation. Since this micro-editing process is only a prelude to selective editing, the mechanism of automatic correction of raw data is quite basic. A common indicator is the growth rate $y_{i,t} \big/ y_{i,t-k}$, an appealing measure based on the time dimension. Except for non-responses which are systematically imputed, only very large outliers are usually corrected at this stage. In fact, the micro-editing process prepares data for selective editing allowing for the computation of relevant aggregates and individual response scores.

5. The second step is the selective editing process, done by usual methods detailed for example in Brion (2007): “drop-out” methods based on score functions measuring the impact of each unit on a given ratio, and “diff” methods using score functions measuring the weighted difference between the raw data and an expected value, an aggregate growth rate for example. Each method is applied on micro-checked data.

6. In short-term statistics, the “diff” score is usually based on the absolute contribution of a business response to the growth rate of an aggregate between dates *t* and *t-k*:

$$c_i = \left| \frac{w_{i,t}y_{i,t} - w_{i,t-k}y_{i,t-k}}{\sum_{j\in A} w_{j,t-k}y_{j,t-k}} \right| = \left| \frac{w_{i,t}y_{i,t} - w_{i,t-k}y_{i,t-k}}{w_{i,t-k}y_{i,t-k}} \times \frac{w_{i,t-k}y_{i,t-k}}{\sum_{j\in A} w_{j,t-k}y_{j,t-k}} \right|$$

In the selective editing process, businesses with a high value of c_i are called back for verification.

7. The “drop-out” score is defined as the growth rate of the aggregate computed with and without the business response:

$$s_i = \left| \frac{\sum_j w_{j,t}y_{j,t}}{\sum_j w_{j,t-k}y_{j,t-k}} - \frac{\sum_{j\neq i} w_{j,t}y_{j,t}}{\sum_{j\neq i} w_{j,t-k}y_{j,t-k}} \right|$$

Once more, only businesses with a high value of s_i are checked.

8. What cut-off values can be used in the selective editing for the *c* and *s* scores? Simulations have been done to find these “gold numbers”. But this is more an interesting theoretical question than a practical one. A short-term indicator must be published roughly at *t*+40, where *t* is the end of the reference period, and you might even have to deliver a flash estimate at *t*+30. Very few raw data are available before *t*+10 and you have little time to collect the data, do the micro-editing, compute a first estimate of the aggregates, do the selective editing, compute the new aggregates, do the seasonal adjustment, perform a new selective editing and compute the final raw and seasonally adjusted indicators. Therefore in practice you do what you can, starting from the top of the list, and you stop when it is time to compute the final estimates, hoping that you did “enough but not too much” and trusting the automatic procedures for correcting what you could not correct on time.

II. Selective Editing Techniques and Seasonal Adjustment

A. The seasonal adjustment process

9. Seasonal adjustment is the last step in the production of Short term economic statistics. It starts just a few days before the release of the indicators. For the Industrial production index (IPI), 1229 seasonally adjusted series are published each month, out of which 613 are directly adjusted and the remaining 616 are derived by aggregation. All these series have to be checked in particular for the absence of residual seasonality and trading-day effects.

10. The series are adjusted using the X-12-ARIMA methodology[1] and the Demetra+ software. All the process is compliant with the “SSE guidelines on Seasonal Adjustment” (Eurostat-ECB, 2009). A “partial concurrent” strategy is used: a very detailed analysis of the series is conducted once a year and the main parameters of the adjustment (ARIMA model, set of trading-day regressors, outliers, length of the filters etc.) are fixed. Each month, the adjustments are done using this set of parameters, re-estimating the coefficients of the Reg-ARIMA model and the components of the series.

11. In a perfect world, the previous editing process should have corrected the problems in the raw data concerning the month under review. And we are here concerned by two issues:

[1] To be simple, X-12-ARIMA adjusts the series in two steps. It first detects and corrects the series for outliers and trading-day effects using a Reg-ARIMA model; the series is also forecasted. In a second step, the method extracts the components of the series (trend-cycle, seasonality, irregular) using a set of predefined moving averages.

- The quality of the seasonally adjusted series;
- The revisions induced by both the seasonal adjustment and the revisions of past raw data.

B. Selective editing

12. X-12-ARIMA outputs a large set of quality statistics, and a few synthetic quality measures, the M and Q statistics (see Ladiray and Quenneville, 2001). The 11 M-statistics measure the size and the randomness of the irregular component (M1 to M6), the presence of seasonality (M7) and the size and evolution of the seasonal component (M8 to M11). The Q-statistic is defined as a linear combination of the M-statistics:

$$Q = \frac{10M1 + 11M2 + 10M3 + 8M4 + 11M5 + 10M6 + 18M7 + 7M8 + 7M9 + 4M10 + 4M11}{100}$$

It is therefore some kind of score function, measuring the quality of the adjustment. Its interpretation is very simple: if $Q \leq 1$ then the adjustment is correct and if Q or several M-statistics are greater than 1, something went wrong during the adjustment.

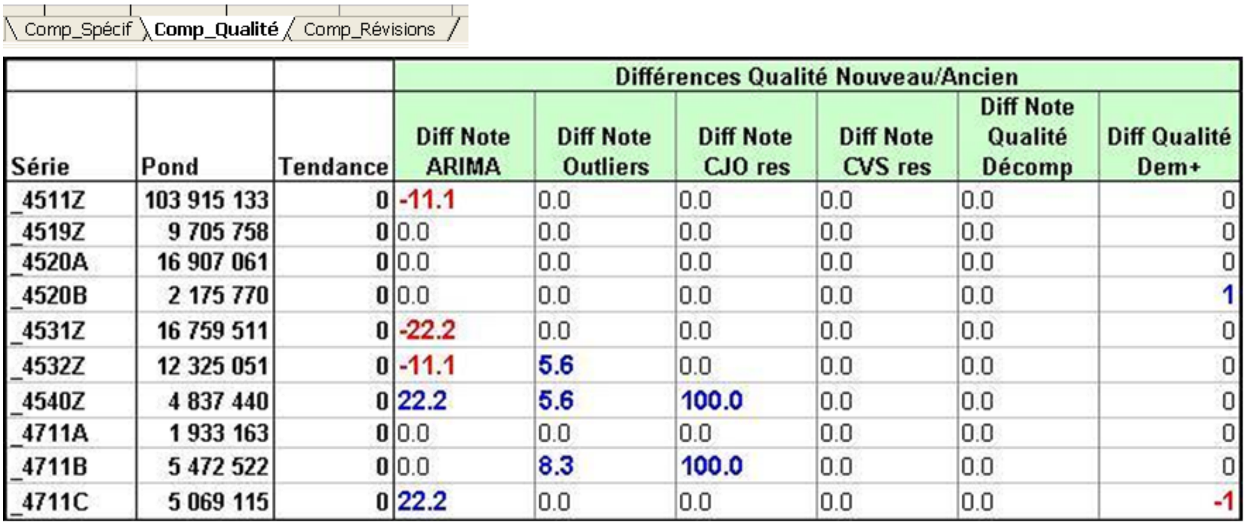
13. In fact, these M and Q-statistics have two defaults. On the one hand, they do not take into account the Reg-ARIMA part of the seasonal adjustment and, on the other hand, they are not computed (at least in Demetra+) for the adjusted series derived by aggregation. A new score function is defined, linear combination of a set of statistics measuring the quality of the adjustment:

$$s_i = f\left(c_{i,1}, \ldots, c_{i,l}, \ldots, c_{i,L}\right)$$

The score is computed for each seasonally adjusted series using statistics for the following aspects of the adjustment:

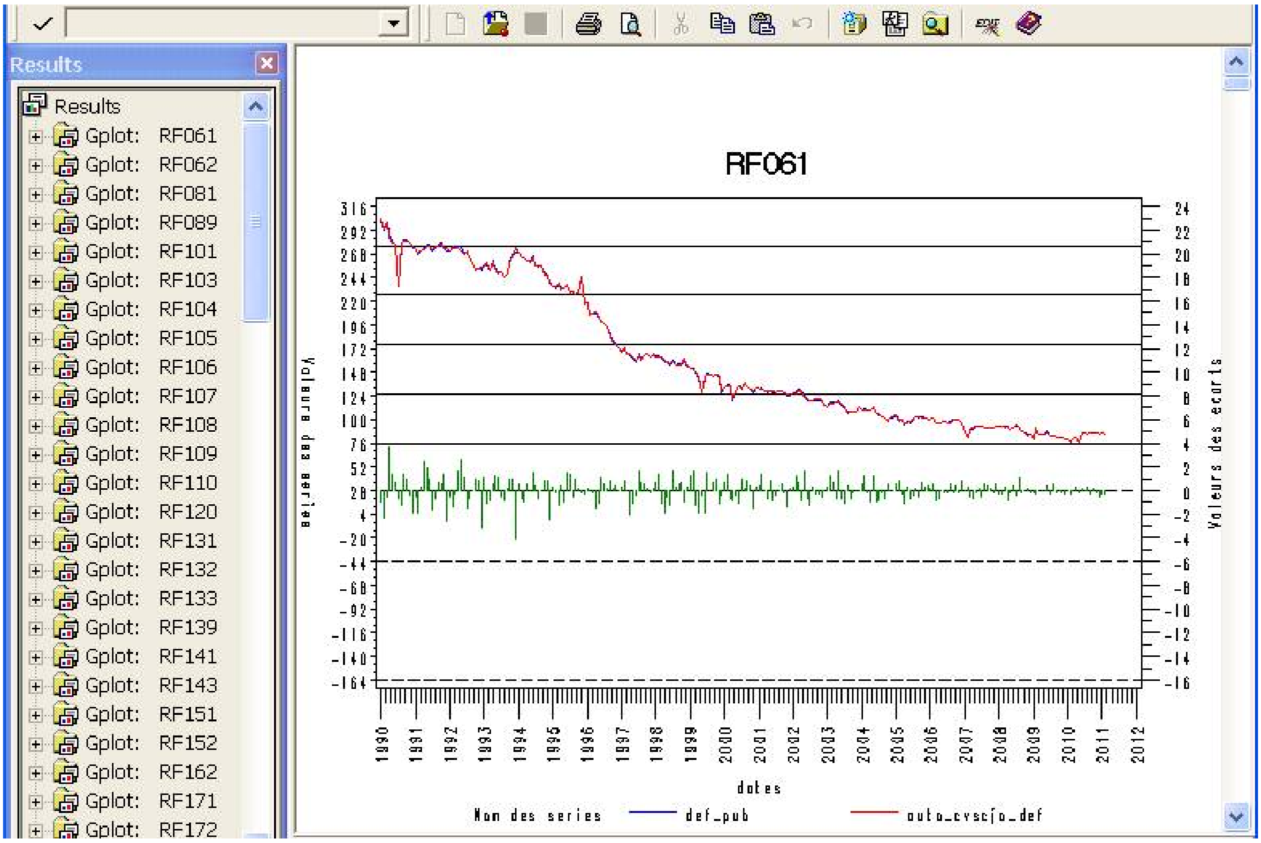
- Number and concentration of outliers;
- Statistical properties of the Reg-ARIMA residuals (lack of correlation, normality, heteroskedasticity, skewness, kurtosis, lack of residual seasonality and trading-day effects etc.);
- Lack of residual seasonality and trading-day effects in the irregular component and in the seasonally adjusted series;
- M and Q-statistics;
- Revision analysis (stability of the model, stability of the components).

14. In production time, the analyst doing the seasonal adjustment receives an Excel file summarizing the quality of the adjustments. Each adjustment is compared to the adjustment done the previous month and colours indicate if the adjustment deteriorates (red) or improves (blue). The results can also be rank according to the weight of the series in the aggregates.



			Différences Qualité Nouveau/Ancien					
Série	Pond	Tendance	Diff Note ARIMA	Diff Note Outliers	Diff Note CJO res	Diff Note CVS res	Diff Note Qualité Décomp	Diff Qualité Dem+
_4511Z	103 915 133	0	-11.1	0.0	0.0	0.0	0.0	0
_4519Z	9 705 758	0	0.0	0.0	0.0	0.0	0.0	0
_4520A	16 907 061	0	0.0	0.0	0.0	0.0	0.0	0
_4520B	2 175 770	0	0.0	0.0	0.0	0.0	0.0	1
_4531Z	16 759 511	0	-22.2	0.0	0.0	0.0	0.0	0
_4532Z	12 325 051	0	-11.1	5.6	0.0	0.0	0.0	0
_4540Z	4 837 440	0	22.2	5.6	100.0	0.0	0.0	0
_4711A	1 933 163	0	0.0	0.0	0.0	0.0	0.0	0
_4711B	5 472 522	0	0.0	8.3	100.0	0.0	0.0	0
_4711C	5 069 115	0	22.2	0.0	0.0	0.0	0.0	-1

15. If something goes wrong, the analyst can see immediately what series, an even what business, is the cause of the problem. He can interactively modify the parameters of the adjustment, re-run Demetra+ and compare the two adjustments.



III. Next Steps

A. Improving the imputation of missing data

16. At the moment the time dimension is not really taken into account when imputing for non-response. In fact we have, at least for the biggest businesses, a quite long time series of their production. As suggested in Aelen (2004), Reg-ARIMA models could be used to obtain an optimal forecast of the series (and of the missing value). A convenient Reg-ARIMA model has to include trading-day regressors and the aggregate computed on available data. A "benchmark series" is necessary because a simple ARIMA modelling is likely to give worst results as it usually cannot forecast turning points.

B. Improving the score function

17. The score function used in the selective editing during the seasonal adjustment process is basic and perhaps too simple. It could be improved in two ways:

- We have to compute more quality statistics for the indirect seasonally adjusted series. M and Q-statistics can easily be computed as well as statistics on the lack of residual seasonality and/or trading-day effects.
- We have to find a better set of weights for the score function, for example using a principal component analysis.

IV. References

[1] Aelen F., 2004, Improving Timeliness of Industrial Short-term Statistics Using Time Series Analysis, Discussion paper N°04005, Statistics Netherlands.

[2] Brion Ph., 2007, "Balance between accuracy and delays in the statistical surveys", INSEE working papers E2007/18, available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/GUIDELINES_FOR_BALANCE_BETWEEN_ACCURACY_AND_DELAYS.pdf

[3] Eurostat-ECB, 2009, ESS Guidelines on Seasonal Adjustment, available at: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-09-006/EN/KS-RA-09-006-EN.PDF

[4] Guggemos F., 2010, « Rapport du groupe de travail sur les procédures de contrôles et redressements dans le cadre du programme Premice », INSEE, Working paper

[5] Ladiray D., Quenneville B., 2001, *Understanding the X11 Seasonal Adjustment Method*, Springer-Verlag, Lecture Notes in Statistics n°158, New York.

[6] Luzi O. et al, 2005, "Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys", available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf