

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Rome, Italy, 2-4 June 1999)

Topic (iii): New methodological and technological developments in statistical data editing

SELECTIVE EDITING METHODS BASED ON TIME SERIES MODELLING.

Submitted by National Statistical Office, Spain¹

Invited paper

Abstract

A selective editing procedure that makes use of ARIMA modelling is presented in this paper. Following the selective editing philosophy, it tries to solve two questions: (1) to characterise and to detect outliers in the macrodata and (2) to define the influential microdata. This procedure is being used in the Spanish National Statistical Institute to elaborate the Industrial Production and Price Indices. Savings in cost and in time of dissemination are being achieved.

KEYWORDS: *continuous surveys, selective data editing, univariate ARIMA models, Intervention Analysis.*

I. INTRODUCTION

1. The central idea of the approach presented in this paper comes from the fact that continuous surveys lead to a set of sequential observations collected over time. Therefore, in these surveys, the appropriate theoretical framework for their study should not be limited to that of the static random variables but should rather be enlarged on random variables varying with time (i.e. the stochastic processes). Indeed, if useful information on previous surveys is available, it should be used to the maximum in different phases of the statistical production process.
2. The use of information of previous surveys is not new in statistical methodology and practice. Ratio and regression estimates or benchmarking techniques are only some examples. In data editing, the use of data from previous surveys is of general application. One of the most frequent ratio edits use the data of the previous survey. Also, monthly, quarterly and annual rates are often used.
3. However, these methods are based on a partial use of the information of the previous surveys. It would be convenient to use, in an efficient way, the whole set of available information, that is, the whole past of the series. This means taking advantage of the whole structure of correlation (cross and auto-correlation). To

¹ Prepared by Pedro Revilla and Pilar Rey.

achieve this, it is necessary to use models that have stochastic processes as a theoretical framework, such as time series analysis models.

4. In this paper, the use of very simple time series models is proposed: univariate ARIMA models (Box-Jenkins, 1970) and univariate ARIMA with Intervention Analysis models (Box - Tiao, 1975).
5. From a theoretical point of view, multivariate models (that picked up the correlation of all the variables) would be appropriate in surveys with more than one variable. However, the difficulty of their practical use suggests the desirability of a univariate environment.
6. ARIMA modelling (in addition to their common use in seasonal adjustment) may be used in statistical offices for data editing and imputation, the description of the data's characteristics for analysis and quality control (Revilla et al., 1991), and linking series. This paper is restricted to the use of ARIMA modelling for data editing.
7. The most useful information for editing short term indicators is the data from previous periods for the same population. For example, monthly and annual rates are often used. Editing based on monthly and annual rates can be improved using ARIMA modelling.
8. The basic idea of this approach is very simple: if the observed data differ considerably from the ARIMA forecast, the data can be erroneous.
9. ARIMA modelling method has the following advantages over the traditional monthly and annual rates:
 - (i) Monthly and annual rates use just one value of previous data. On the contrary, the ARIMA forecast uses the whole set of previous data in an optimal way. In fact, the ARIMA forecast is a linear function of the latter.
 - (ii) The ARIMA forecast enables us to use probabilistic data editing. This allows to take into account the different variability of the economic sectors, products, etc.

II. DESCRIPTION OF THE SELECTIVE EDITING PROCEDURE

10. The approach presented here is being used in the Spanish National Institute to elaborate the Industrial Production and Price Indices. In this paper we concentrate on the Industrial Production Indices. Similar formulas are used for the Price Indices. It could also be implemented for other short-term indicators.

11. A monthly survey is carried out by mail in order to calculate the Industrial Production Indices. A panel sample of about 9000 enterprises is used. The response rate is about 95%.

12. One single variable, the production volume, measured in physical units (tons, litres, etc.) or in monetary value, is requested from each enterprise. As a result of the survey, we have a microdata set with elements $q_{i,j,t}$, that is, the production figure for the product i , reported by the enterprise j at month t .

13. From the microdata set, the index for product i is calculated as:

$$I_{i,t} = I_{i,t-1} \frac{\sum_j q_{i,j,t}}{\sum_j q_{i,j,t-1}}$$

where j is the set of enterprises with valid values at both t and $t-1$.

14. From these product indices, Laspeyres aggregated indices are calculated at successive levels of breakdown of the economic activities classification (at the top of the aggregation is the total industry). The following formula is used:

$$I_t = \sum_i w_i I_{i,t}$$

where the base year weights w_i are based on the value added (for activities aggregation) or the value of the production (for products aggregation).

15. A traditional approach was used for data editing, consisting of the following steps:

- (i) microediting, mainly using monthly and annual rates;
- (ii) indices computation;
- (iii) indices macroediting, using monthly and annual rates and subject matter judgement about the behaviour of each of the series;
- (iv) and, again, microediting the individual data of the indices that are suspicious of error.

16. The former process was carried out in an iterative way until all the indices were considered valid.

17. The disadvantages of that approach were:

- a low "hit rate" (ratio of editing changes to the number of flags) was achieved;
- the same microdata were revised many times;
- identical efforts were made to edit microdata with great and small impacts on macrodata;
- the editing criteria were often subjective.

18. We have tried to solve or to minimise these problems using a selective editing strategy, with the following targets: to improve the "hit rate", to integrate the editing phases, to prioritise the efforts on errors with a great impact on the macrodata, and to use objective criteria.

19. Following the selective editing philosophy, the procedure tries to solve two problems:

- (i) to define and to detect outliers in the macrodata (the indices);
- (ii) to define and to detect the influential microdata.

20. In order to achieve the first target we have designed some tools, the "surprises", that are functions of the ARIMA model forecast.

21. Since the number of time series to handle is very large and it is difficult and time consuming to build models for all of them we need an automatic procedure. We use an automatic method developed by Revilla, Rey and Espasa that fits into the Box-Jenkins iterative modelling strategy of identify, estimate and diagnostic checking. Using this method, an ARIMA model has been constructed for each of the index series of products and activities.

22. A straightforward use of ARIMA models is not sufficient to capture calendar variations in the indices, because they are not exactly periodic. Regression models are used to handle calendar effects and other deterministic variations (for example, a strike). To specify the intervention variables we have found that some subject matter knowledge about the behaviour of the indices is needed. Therefore, the overall models are a sum of ARIMA and regression models:

$$\ln I_{i,t} = \frac{q_i(B) \Theta_i(B^{12})}{j_i(B) \Phi_i(B^{12})} a_{i,t} + \sum_h \frac{a_{i,h}(B)}{d_{i,h}(B)} A_{i,h,t}$$

where:

- $\ln I_{i,t}$ is the neperian logarithm of the industrial production index for product (or activity) i .
- B is the backshift operator, $B^k(I_t) = I_{t-k}$.
- $q_i(B), j_i(B), \Theta_i(B^{12}), \Phi_i(B^{12}), a_{i,h}(B), d_{i,h}(B)$ are polynomials in the backshift operator.
- $a_{i,t}$ are white noise variables i. i. d. $N(0, S)$.
- $A_{i,h,t}$ are intervention variables.

23. If we calculate the one-step ahead forecast $\ln \hat{I}_{i,t}$ for $\ln I_{i,t}$, the one-step ahead forecast error is:

$$e_{i,t} = \ln I_{i,t} - \ln \hat{I}_{i,t}$$

24. From these models (and, in particular, from the one-step ahead forecasted values) we can construct the following tools:

25. The **Surprise (or simple surprise)** $S_{i,t}$ for the index $I_{i,t}$ is the relative change between the observed and the forecasted data:

$$S_{i,t} = \frac{I_{i,t} - \hat{I}_{i,t}}{\hat{I}_{i,t}}$$

26. Since the one-step ahead forecast error $e_{i,t}$ is a $N(0, S_i)$ white noise process and $\ln I_{i,t} - \ln \hat{I}_{i,t} \cong (I_{i,t} - \hat{I}_{i,t}) / \hat{I}_{i,t}$, we have that $S_{i,t}$ is approximately $N(0, S_i)$. Hence, a confidence interval (for example, a 95% interval) for the surprises can be constructed:

$$P[-1.96 S_i < S_{i,t} \leq 1.96 S_i] = 0.95$$

and the outliers can be defined as the indices whose surprise is outside the interval.

27. The **Standard surprise** for the index $I_{i,t}$ is:

$$\frac{S_{i,t}}{S_i} = \frac{I_{i,t} - \hat{I}_{i,t}}{\hat{I}_{i,t}} \frac{1}{S_i}$$

It allows the direct comparison of indices with different variability.

28. The *Weighted standard surprise* for the index $I_{i,t}$ is:

$$\frac{S_{i,t}}{S_i} w_i = \frac{I_{i,t} - \hat{I}_{i,t}}{\hat{I}_{i,t}} \frac{w_i}{S_i}$$

It allows the ranking of the indices taking into account not only the surprise magnitude but also the different weights.

29. Once we have detected and ranked the surprising indices (i.e., indices that are not coherent with their past behaviour and therefore can be considered as outliers) we need to measure the impact of each of the microdata on these surprising indices. For this purpose, we use the “influences”.

30. The *Influence of an individual datum over an aggregated magnitude* is defined as the difference between the observed aggregated magnitude and the value for this same magnitude when the individual datum is not available.

31. The *Influence of the individual datum $q_{i_0,j_0,t}$ over the product index $I_{i_0,t}$* is:

$$INF_{i_0,j_0}^{I_{i_0,t}} = I_{i_0,t-1} \frac{\sum_j q_{i_0,j,t}}{\sum_j q_{i_0,j,t-1}} - I_{i_0,t-1} \frac{\sum_{j \neq j_0} q_{i_0,j,t} + \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}} = I_{i_0,t-1} \frac{q_{i_0,j_0,t} - \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}}$$

where $\hat{q}_{i_0,j_0,t}$ is an imputed value for the individual datum $q_{i_0,j_0,t}$.

32. The *Influence over the aggregated index I_t* is:

$$INF_{i_0,j_0}^{I_t} = \sum_i w_i I_{i,t} - \left[\sum_{i \neq i_0} w_i I_{i,t} + w_{i_0} I_{i_0,t-1} \frac{\sum_{j \neq j_0} q_{i_0,j,t} + \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}} \right] = w_{i_0} I_{i_0,t-1} \frac{q_{i_0,j_0,t} - \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}}$$

33. This expression measures the impact of the microdata on the index by means of the following factors:

- The product (or activity) weight w_{i_0} .
- The index $I_{i_0,t-1}$ which “updates” the above weight.
- A measure of the relative discrepancy between the real and the imputed individual datum $\frac{q_{i_0,j_0,t} - \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}}$.

34. These “influences” allow us to prioritize the suspicious values in the microdata in order to verify and recontact fewer enterprises. It may be proved that the microdata which are more influential on the aggregated index are also the more influential on the surprises of that index.

III. FINAL REMARKS

35. This selective editing procedure fits into the Total Quality Management strategy implemented in the Spanish Industrial Surveys: the editing process is now much more integrated, the repetition of some stages has been eliminated and many tedious tasks have been replaced by fewer and more qualified ones.

36. And, what is more important, there have been improvements in timeliness (the first quality requirement for our customers) simultaneously with reductions in the resources needed (fewer hours of work for the editing tasks) and in the response burden (fewer recontacts).

REFERENCES

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

Box, G.E.P. and Tiao, G.C. (1975). *“Intervention Analysis with Applications to Economic and Environmental Problems”*, Journal of the American Statistical Association, 349.

Granquist, L. (1995). *“Improving the Traditional Editing Process”*, Business Survey Methods. Wiley Series in Probability and Mathematical Statistics.

Revilla, P., Rey, P. and Espasa, A. (1991). *“Characterisation of Production in Different Branches of Spanish Industrial Activity by means of Time Series Analysis”*. Working Papers. Carlos III University. Madrid.