STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 13
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Data Editing
(27 – 29 May 2002, Helsinki, Finland)

Topic (ii): Measuring and evaluating data editing quality

AN E&I METHOD BASED ON TIME SERIES MODELLING DESIGNED TO IMPROVE TIMELINESS

Invited paper

Submitted by the National Statistical Institute, Spain[1]

Abstract

An edit and imputation method, based on time series modelling, is presented in this paper. The method has been designed to improve timeliness, nowadays an essential quality requirement of public statistics (in particular of short-term indicators), maintaining the current levels of accuracy. REGARIMA modelling is used as a tool to carry out both editing and imputation. The method is being implemented to produce Spanish monthly short-term indicators. An example of using this method in the Industrial Production Indices is also presented.

KEYWORDS: short- term indicator, TQM, selective data editing, REGARIMA modelling

I. INTRODUCTION

1. This paper follows the approach already presented in previous meetings (i.e. using time series modelling for data editing). In Rome we presented a selective editing procedure based on a kind of tools that we named “surprises”, which are functions of ARIMA forecast. In Cardiff, we used time series modelling in a different way: we obtain a set of characteristics from an estimated ARIMA with Intervention Analysis model in order to use them to improve the edits. For this meeting, the focus is on using time series modelling to improve timeliness.

2. Nowadays, public statistical offices are under continuous pressure from society, which demands more and more data, to be produced at a lower cost, with a lower respondent burden, and especially with a shorter delay. Achieving timeliness though with losses of accuracy, by trying to solve the problem in the short term, may be tempting. But it would not be a very good strategy, because public confidence in statistics would probably suffer. Therefore, the only suitable way is trying to reduce dissemination time, without any losses in accuracy.

3. Timeliness and accuracy are usually considered as a major trade-off in the production of statistics. Nevertheless, new IT tools and statistical methodologies offer the opportunity for re-engineering statistical production processes in order to improve timeliness and accuracy simultaneously.

4. Traditionally, data editing is linked with accuracy, but increasingly is also linked with other quality aspects (Kovar, 1997). In this context, some crucial questions arise. For example, what

[1] Prepared by Pedro Revilla (previlla@ine.es).

role should data editing play in achieving timeliness? Or, how can we redesign data editing processes in order to improve timeliness? Or, which new techniques can we introduce to help us to get this target?

5. This paper discusses some of these questions, presenting our experiences in improving timeliness. An edit and imputation method based on time series modeling, in particular, on REGARIMA modeling (Bell, 1999), is a major methodological tool for improving timeliness in short-term indicators. It fits into a more general programme carried out for all the Spanish industrial statistics. The target of the program is to reduce the dissemination dates maintaining the level of quality. The implementation of the programme is a direct consequence of the TQM approach used to produce industrial statistics.

6. In the following section, a general description of the TQM approach and the timeliness programme used is presented. In section III, the use of time series modeling for editing short-term indicators is discussed. An application for the Spanish Industrial Production Indices is presented in section IV. The paper ends with some conclusions.

II. TQM APPROACH AND TIMELINESS PROGRAM USED FOR INDUSTRIAL STATISTICS

7. Trying to follow TQM principles (in particular, customer satisfaction, cornerstone of TQM), the approach used in the industrial statistics (Gonzalez and Revilla, 1997) is quite simple. The starting point is asking customers about our main failures. Then, specific programs and actions are launched to correct those failures and to improve customer satisfaction.

8. Based on our customers' opinions, we have learned that the main failure of our statistics is the delay in the data production. Eighty percent of customers consider the delay our main failure. Therefore, it should constitute an essential priority to improve our timeliness.

9. To face the problem of timeliness we have launched a general programme for the industrial statistics in order to improve timeliness, maintaining the level of accuracy. Adapting questionnaires to the accounting practices of the enterprises and improving our relationship with enterprises are two key factors in achieving timeliness. Introducing selective editing methods is probably the main technical factor.

10. We have taken several measures in order to adapt questionnaires to the accounting practices of the reporting enterprises (adapting variables and valuation rules, using different models of questionnaires, using personalized questionnaires, etc.). The underlying principle in this approach is that enterprises provide data in the same way they produce them for their own use, and the statistical agency re-elaborates them for analytical purposes, if necessary. Adapting questionnaires to the accounting practices makes answering them easier and quicker. It also results in fewer errors made by the enterprises. All of this leads to an improvement both in accuracy and time of dissemination.

11. Another key success factor in achieving timeliness is improving our relationship with reporting enterprises (our suppliers). We offer tailored data on market share, in exchange for the questionnaires. Hence we change our suppliers into customers. We are trying to work using a new model of relationship with enterprises, that we name the joint venture model. The underlying principle is that our relationship with reporting enterprises should be based on a relationship of mutual use and collaboration, rather than on the legal duty of the enterprises to fill in compulsory questionnaires. Because of the increasing interest of enterprises, they are filling in

questionnaires sooner and with more care. An important point of using the joint venture model is that both timeliness and accuracy can be improved, reducing the perceived enterprise burden at the same time.

12. Timeliness should be considered a major goal in our production rather than just a small problem. Therefore, we need to save time in all the phases of the statistical process. Data editing is one of the most time-consuming statistical phases. Hence, re-engineering the data editing procedures is a need for improving timeliness. The use of selective editing methods is a key factor to improve timeliness. Using time series modelling joined to the selective editing philosophy is being quite useful to save time in editing short-term indicators.

III. THE USE OF TIME SERIES MODELLING FOR EDITING SHORT-TERM INDICATORS

13. Time series modelling is not commonly used in statistical agencies for producing short-term indicators, with the exception of seasonal adjustment.

14. Nevertheless, 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 random variables but should rather be enlarged on random variables varying with time (i.e. the stochastic processes). Therefore, the use of models that have stochastic processes as a theoretical framework (such as time series analysis models) may be very useful. Indeed, if useful information on previous surveys is available, it should be used to the maximum in different phases of the statistical production process.

15. Certainly, the use of information of previous surveys is not new in data editing. One of the most frequent ratio edits is based on the data of the previous survey, and monthly, quarterly and annual rates are of general application. 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 entire past of the series. This means taking advantage of the whole structure of correlation (cross and auto-correlation). Another advantage of using time series modelling is that it enables us to use probabilistic data editing. This is very useful because it allows taking into account the different variability of the economic sectors, products, etc.

16. Time series modelling can be used adopting very different editing strategies. We can either build models for the microdata (or a subset of them, for example the biggest enterprises) for editing microdata and hence use a microediting approach, or we can build models for the macrodata and use a macroediting or a selective editing approach. For a specific survey the best approach would probably be combining all of them in order to improve timeliness.

IV. AN APPLICATION FOR THE INDUSTRIAL PRODUCTION INDICES

17. An edit and imputation method based on REGARIMA modelling is being used in the Spanish National Statistical Institute to elaborate industrial short-term indicators. The models are used for microediting, macroediting and selective editing. Micro and macro imputation are also carried out based on the models. In this paper we focus on the Industrial Production Indices. Similar formulas are used for other short-term indicators.

18. A monthly survey is carried out by mail in order to calculate the Industrial Production Indices. A panel sample of about 14000 enterprises is used. One single variable, the production, in a particular physical unit (tons, litres, etc.) or in monetary value, is requested from each

enterprise. As a result of the survey, we have a microdata set $q_{i,\,j,\,t}$, that is, the production figure for the product i, reported by the enterprise *j* at month t. 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 $t-1$.

And, 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).

19. We use the same kind of models (REGARIMA) for the micro and the macrodata series. 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 (Box and Jenkins, 1970). A straightforward use of ARIMA models is not sufficient to capture calendar variations, 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 production data is needed.

20. Therefore, the overall models are a sum of ARIMA and regression models (REGARIMA models):

$$ln\, q_{i,j,t} = \frac{\theta_{i,j}(B)\,\Theta_{i,j}\left(B^{12}\right)}{\varphi_{i,j}(B)\,\Phi_{i,j}\left(B^{12}\right)} a_{i,j,t} + \sum_{h} \frac{\alpha_{i,j,h}(B)}{\delta_{i,j,h}(B)} A_{i,j,h,t} \quad \text{for modelling the microdata.}$$

$$ln\, I_{i,t} = \frac{\theta_{i}(B)\,\Theta_{i}\left(B^{12}\right)}{\varphi_{i}(B)\,\Phi_{i}\left(B^{12}\right)} a_{i,t} + \sum_{h} \frac{\alpha_{i,h}(B)}{\delta_{i,h}(B)} A_{i,h,t} \quad \text{for modelling the indices.}$$

where:

- $ln\, q_{i,j,t}$ is the neperian logarithm of the production figure for the product i, reported by the enterprise *j*.
- $ln\, I_{i,t}$ is the neperian logarithm of the industrial production index for product (or activity) i.
- B is the backshift operator, $B^k\left(I_t\right) = I_{t-k}$.
- $\theta\ (B), \varphi\ (B), \Theta\ \left(B^{12}\right), \Phi\ \left(B^{12}\right), \alpha_h(B), \delta_h(B)$ are polynomials in the backshift operator, for the product (or activity) i, or for the enterprises i,j.

- $a_{i,j,t}$ and $a_{i,t}$ are white noise variables i. i. d. $N(0,\sigma_{i,j})$ and $N(0,\sigma_i)$ respectively.
- $A_{i,j,h,t}$ and $A_{i,h,t}$ are intervention variables.

21. For using a **microediting approach** we would have to fit a REGARIMA model for each series of enterprise data. In our application, we do not fit a model for all of the enterprises but only for the most influential ones (usually the biggest enterprises). The model is used for both editing and imputation. A confidence interval (for example a 95% interval) can be constructed from the model:

$$P\left[\hat{q}_{i,j,t} - 1{,}96\ \sigma_{ij} < q_{i,j,t} < \hat{q}_{i,j,t} + 1{,}96\ \sigma_{i,j}\right] = 0{,}95$$

Where $\hat{q}_{i,j,t}$ is the one-step forecast for $q_{i,j,t}$, and the outliers can be defined as the microdata outside the interval. The imputed value for $q_{i,j,t}$ would be $\hat{q}_{i,j,t}$

22. For using a **macroediting approach**, an REGARIMA model has been constructed for each of the index series of products and activities. The model is used for both editing and imputation. A confidence interval (for example a 95% interval) can be constructed from the model:

$$P\left[\hat{I}_{i,t} - 1{,}96\ \sigma_i < I_{i,t} < \hat{I}_{i,t} + 1{,}96\ \sigma_i\right] = 0{,}95$$

Where $\hat{I}_{i,t}$ is the one-step forecast for $I_{i,t}$, and the outliers can be defined as the indices outside the interval. The imputed value for $I_{i,t}$ would be $\hat{I}_{i,t}$

23. In using a **selective editing approach** we have to solve two problems: to detect outliers in the macrodata (the indices) and to detect the influential microdata. In order to face the first problem we have designed some tools, the “surprises”, that are functions of the REGARIMA model forecast (in particular, from the one-step ahead forecasted values):

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:

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

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:

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

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

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

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

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

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

It allows the direct comparison of indices with different variability.

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

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

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

24. 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".

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.

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 $q_{i_0,j_0,t}$.

and the ***Influence over the aggregated index*** I_t is:

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

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 $\dfrac{q_{i_0,j_0,t} - \hat{q}_{i_0,j_0,t}}{\sum_j q_{i_0,j,t-1}}$.

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.

25. These "influences" allow us to prioritize the suspicious values in the microdata in order to verify and recontact fewer enterprises. Hence improvements in timeliness are achieved.

26. As an example of using this methodology two computer printouts are shown. In table 1 the sectors are ranked according the weighted standard surprise. Sector 4243 is the one with the highest. Looking in table 2 for the enterprises that most influence the index of this sector, we find an enterprise whose influence is not only the highest, but also much higher in comparison with the others. Hence this enterprise should be checked.

27. Using this methodology, the most influential suspicious values could immediately be detected. Hence the most important errors could be checked and corrected before the index is disseminated for first time.

28. Three kinds of **imputations** would be possible to handle the non-response for one or more enterprise data $q_{i,j_o,t}$ to disseminate the indices $I_{i,t}$:

(i) A traditional imputation based on the available enterprise data

$$\hat{q}_{i,j_0,t} = q_{i,j_0,t-1} \frac{\sum_{j \neq j_o} q_{i,j,t}}{\sum_{j \neq j_o} q_{i,j,t}}$$

(ii) A model-based imputation for $q_{i,j_o,t}$ based on its past series, that is $\hat{q}_{i,j_o,t}$

(iii) A model-based imputation for the whole index $I_{i,t}$ based on its past series, that is $\hat{I}_{i,t}$

29. A choice among the former methods can be made depending on various circumstances. For example, if the trend of the enterprises inside the product (or activity) is similar, the traditional imputation usually works properly. However, if the trend is not very similar a model-based imputation works better. In the same way, if the non-response rate is low for the index, a micro imputation (traditional or model-based) usually works properly. On the contrary, if the non-response rate for the index is high and there are not many enterprises in the index a model-based imputation for the whole index is preferable. Therefore, model-based imputation may improve the estimates of the indices.

V. CONCLUSIONS

30. Improving timeliness without any losses in accuracy is a major challenge for public statisticians today. Data editing should not only be linked to accuracy but also to other quality

aspects, for example timeliness. Data editing is one of the most time-consuming statistical phases. Hence, re-engineering the data editing procedures is a need for improving timeliness.

31. According to our experience, the use of time series modeling is being quite useful to save time in editing short-term indicators.

REFERENCES

[1] Bell, W.R. (1999). *"An overview of REGARIMA modeling*". Forthcoming Research Report. Statistical Research Division. U.S. Census Bureau.

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

[3] González, M. and Revilla, P. (1997). **"***Total Quality Management and the INE"*. Eurostat Web, www.forum.europa.eu.int/.

[4] Kovar, J. (1997). *"What to do when an edit fails*". Statistical data editing. Volume No.2. Conference of European Statisticians. Statistical standards and studies-No.48.

[5] Revilla, P. and Rey, P. (1999*).* **"***Selective editing methods based on time series modelling"*. UN/ECE Work Session on Statistical Data Editing. Rome.

[6] Rey, P. and Revilla. P, (2000). **"***Analysis and quality control from ARIMA modelling"*. UN/ECE Work Session on Statistical Data Editing. Cardiff.

Table 1 SURPRISES

Sector	Actual Rate	Forecasted Rate	Simple Surprise	Standard Surprise	Weighted Standard Surprise
4243	70,28	3,32	64,73	3,79	17,10
2511	-27,73	-3,29	-25,25	-3,11	-16,93
4110	-50,24	-6,89	-70,96	-6,84	-16,89
2514	-15,92	4,64	-19,62	-3,00	-16,87
2512	39,39	-11,83	58,12	7,22	16,51
4752	-0,74	2,06	-2,75	-1,09	-15,66
3299	-11,97	4,45	-15,70	-2,02	-15,57
4751	22,82	-7,36	32,55	2,34	14,64
3630	-0,28	3,68	-3,82	-0,81	-14,54
3166	15,97	5,83	9,58	1,89	13,92

Table 2 INFLUENCES

Enterprise	Influence
1	143,41
2	37,60
3	-22,80
4	14,38
5	8,90
6	-7,52
7	7,35
8	6,81