STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 19
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Cardiff, United Kingdom, 18-20 October 2000)

Topic III: New techniques and tools for editing imputation

ANALYSIS AND QUALITY CONTROL FROM ARIMA MODELLING

Submitted by the National Statistical Institute and Sociological Research Centre, Spain[1]

Invited paper

Abstract: In this paper, we use ARIMA modelling to estimate a set of characteristics of a short-term indicator (for example, the index of industrial production), as trends, seasonal variations, cyclical oscillations, unpredictability, deterministic effects (as a strike), etc. Thus for each sector and product (more than 1000), we construct a vector of values corresponding to the above-mentioned characteristics, that can be used for data editing.

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

I. INTRODUCTION

1. This paper follows the approach already presented in the last meeting in Rome (i.e. the use of time series analysis for data editing). The central idea of this approach is that, if useful information from previous surveys is available, it should be used as much as possible in the editing process.

2. In this context, the appropriate theoretical framework for editing continuous surveys should not be limited to that of the static random variables, but should rather be enlarged on that of sequences of random variables varying with time. Therefore, stochastic processes and, in particular, time series models may be used.

3. At the previous meeting in Rome, we presented a selective editing procedure based on a kind of tools that we named “surprises”, which are functions of ARIMA forecasts. For this meeting, we use time series modelling in a different way. We propose to estimate a univariate ARIMA with Intervention Analysis model (Box-Tiao, 1975) to describe the characteristics of a short time indicator.

4. The idea comes from the fact that in editing a survey, we need as much information as possible about the phenomenon that we try to measure. On the other hand, different subsets of data often show a very different variability and behaviour. For example, when editing the Spanish monthly indices of industrial production, we can find extremely small (even zero) production data values for August, because summer holidays are usually taken in this month in Spain. Of course, these data must not be considered as outliers (i.e. suspicious data items) if we have information about this seasonal pattern. But an additional problem is that, this seasonal behaviour in August is very different from one branch to another (even there are branches which production does not decrease but strongly grows in August, as in beer production). For this reason, it is important to acquire information about the different dynamic characteristics of each of the branches to improve editing rules and strategies.

[1] Prepared by Pedro Revilla and Pilar Rey.

5. In this paper, we use ARIMA with Intervention Analysis modelling to estimate a set of characteristics of a short term indicator, as trends, cyclical oscillations, seasonal variations, calendar effects, special event effects (as a strike), unpredictability, etc. Thus for each of the branches and products (more than 1000), we construct a vector of values corresponding to the characteristics above mentioned, that can be used as a useful tool by the data editing team.

6. Even when, from a theoretical point of view, multivariate models (that picked up the correlation of all the variables of the survey) would be appropriate, we restrict ourselves to the univariate environment, because of the difficulties in using multivariate time series modelling in practice. On the other hand, the use of univariate ARIMA models to describe the data characteristics of an economic phenomenon has a sound methodological foundation.

7. Under fairly general conditions (see Prothero and Wallis (1976), Wallis (1977), Zellner (1979), etc.) any variable, which is determined within a structural simultaneous and dynamic econometric model (SEM) is generated in a univariate way by an ARIMA with Intervention Analysis model. In the latter model, the intervention analysis component picks up the contribution of the dummy variables of the SEM model and/or the effect of certain intervention analysis, which affect the exogenous variables of that model.

8. To the extent that the SEM model reflects the characteristics of the real world, the ARIMA with Intervention Analysis model corresponding to an endogenous variable of the SEM model incorporates, inefficiently but certainly consistently, the basic characteristics of that variable.

II. APPLICATION TO THE INDUSTRIAL PRODUCTION INDICES

9. The approach presented here is being used in the National Statistical Institute of Spain for editing the Industrial Production Indices. It could also be implemented for other short-term indicators. 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%.

10. One single variable, the production volume, measured in physical units (tons, litres, etc.) or in monetary value, is requested from each enterprise. The indices for products are calculated as chain indices. From these product indices, Laspeyres aggregated indices are calculated at successive levels of breakdown of the economic branches 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 calculated using the value added (for branches aggregation) or the value of the production (for products aggregation). The branches have been studied at five successive levels of breakdown, in accordance with the National Classification of Economic Activities existing in Spain.

11. In all of the series the sample contains 96 observations, from January 1992 to December 1999. To characterise the evolution of production in the different branches of industrial activity ARIMA models with Intervention Analysis (ARIMA-IA) are used.

12. 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 (1990), 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 series of indices.

13. A straightforward use of ARIMA models is not sufficient to capture calendar effects and other deterministic variations (for example, a strike). Regression models are used to handle them. Therefore, the overall models are a sum of ARIMA and regression models:

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

where:

- $\ln I_{i,\,t}$ is the neperian logarithm of the industrial production index for product (or activity) i.
- B is the backshift operator, $B^k\,(I_t) = I_{t-k}$.
- $q_i\,(B), j_{\,i}\,(B), \Theta_i\left(B^{12}\right), \Phi_i\left(B^{12}\right), a_{i,h}\,(B), d_{\,i,h}(B)$ 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.

14. To specify the intervention variables we have found that automatic procedures are not suitable for all of the series and some subject matter knowledge about the behaviour of the indices is needed.

15. In the following, we will consider one by one the different aspects we have studied in the models, describing how they have been implemented.

A. Level behaviour

16. An adequate description of the nature of the long-term trend of the series is contained in the models. This trend is determined by the contribution in the final forecast function of the real positive unit roots of the autoregresive factor and by the contribution of the possible non-zero mean of the stationary series. The presence of d roots of the type already mentioned means that the long-term trend is a time polynomial of order (d-1), the coefficients of which are determined by the initial conditions in which the system is located. The presence of a non-zero means increases the previous polynomial with a term of order d with a deterministic coefficient.

17. Thus, whenever the series is stationary, it would not require differences. When the model specifies one difference the series will show local oscillations in level; when the model specifies two differences the series will have a quasi-linear trend, etc.

B. Seasonal behaviour

18. The models also can contain a factor, which picks up a seasonal cycle with a period of 12 time units. The complex unit roots and real negatives of the autoregressive polynomial form this factor. If none of these roots is repeated its contribution in the final forecast function consists of 12 stable, additive, seasonal factors which at all times are determined by the initial conditions of the system. As a result, the long-term path of the series is made up of these seasonal factors and the trend described in A.

Whenever a seasonal cycle has been found the seasonal factors of the final forecast function have been calculated.

C. Calendar effects

19. Spanish Indices of Industrial Production series are strongly affected by calendar effects. Because these data are reported monthly and represent a total of the production for each month, they contain variations due to the length and day-of-the-week composition of the month. Monthly production data are also affected by holidays, in which the levels of production are lower than on ordinary working days. There are holidays that occur each year at the same time, and movable holidays, for example Easter, that occur at various dates during March and April. To make the situation more complicated, public holidays in Spain vary from one year to another, and from one "Autonomous Community" to another. In order to handle these calendar effects, we incorporate this information as deterministic input variables.

20. Instead of the most commonly used (see Hillmer, Bell and Tiao, 1983) seven trading-day variables (number of Mondays, Tuesdays, etc.), we construct one single variable, adapted to the behaviour of the Spanish industrial production. This variable tries to measure the number of working days, removing Saturdays, Sundays and holidays. To achieve a suitable computation, those holidays that vary from one "Autonomous Community" to another have to be weighted by the industrial value added of the base year.

21. By constructing the model on the logarithmic transformation of the series, the parameter associated with this dummy variable can be interpreted as the proportional increase in production in comparison with that of a similar month with one less working day.

22. Another deterministic variable containing the Easter effect (which takes in March and April the values that indicate the proportion of days affected by these holidays and in the other months value zero) is included in the model (see Hillmer, Bell and Tiao, 1983). The parameter, which affects this artificial variable, can be interpreted as the proportional variation suffered by production as a result of this effect. Hence, all calendar effects are summarised using only two variables.

D. Other deterministic effects

23. It is possible to find deterministic contributions in the trend and/or seasonal factors of the series from the intervention analysis. In February 1997 the transporting sector went on strike. For most of the industrial branches, this strike caused a shortage of raw materials. The expected effect would have been an immediate reduction in the level of production. But, in fact, after some periods of observation, we found in some series an increase in the months of March and April that compensated the decrease in February. After consulting some respondent factories, we learned that they have tried to fulfil the orders of the clients by working extra in the next two months.

24. So, we have constructed two different variables for peaking up the effects of the strike:

1) The H_t variable, where:

$$H_t = \begin{cases} 1.0, & t = \mathit{February}\ 1997 \\ -0.6, & t = \mathit{March}\ 1997 \\ -0.4, & t = \mathit{April}\ 1997 \\ 0.0, & t \neq \mathit{Feb., Mar.\ orApr.}\ 1997 \end{cases}$$

2) The P_t pulse variable, where:

$$P_t = \begin{cases} 1, & t = February\ 1997 \\ 0, & t \neq February\ 1997 \end{cases}$$

25. We have accepted that the strike had an effect on an industrial production index series when the intervention variable parameter is significantly different from zero. When the parameters are significant for the two of them, we have chosen the intervention variable that produces a smaller residual standard deviation of the ARIMA-IA model. As we have used the logarithmic transformation of the series, it is possible to interpret the value of the parameters as a percentage effect on the level of the original series.

E. Outliers

26. We have made in the series a detection of unusual or unexpected observations. Depending on their nature, we have detected three types of outliers: Additive Outlier (A. O), Level Shift (L. S.), and Temporary Change (T. C.).

27. For defining and detecting the location and type of an outlier, we follow the approach of Chang et al. (1988) using the estimated residuals of the Arima model.

> *The study of outliers can be used to detect special events which may affect production in a particular period of time, to analyse if they appear fortuitously or not in the different branches, the month in which they often appear, etc.*

F. Uncertainty

28. The last characteristic we have considered is a measure of the uncertainty about the future evolution of the series, expressed by the standard deviation of the one period ahead forecasting errors. Each series at a given time can be broken down into two components: its prediction based on previous observations and a prediction error. A good measure of uncertainty is given by the standard deviation of these prediction errors.

III. RESULTS OF THE STUDY

29. As an example of the information that may be obtained from the models, the main dynamic characteristics of the indices for the total industry and for the branches at two digits level are presented in Table 1.

30. The following are shown in the table: 1) Identification of the branch by means of the code of the National Classification of Economic Activities of Spain, 2) Behaviour of the level, 3) Whether the sector has seasonal behaviour or not, 4**)** A measure of the effect of working days, when it affects the series, 5) A measure of the effect of Easter, when it affects the series, 6) A measure of the effect of the 1997 strike, when it affected the series, 7) The outliers, the date when they are produced and their type, 8) The degree of uncertainty about future production levels.

31. To illustrate the information that can be obtained from Table 1, the characteristics of total industry are described. Industrial production from 1992 onwards shows a trend in its level and a stochastic seasonal nature, which implies different behaviour in production activity in different months of the year (in particular, it shows a decrease in the holiday months, especially in August).

32. The industrial output is sensitive to day-of-the-week composition and periods of holidays. More specifically, the existence of one less working day causes a 1,9% drop in production. Likewise, Easter causes a fall in production of 4,0% distributed in March and April according to the proportion of days affected by this holiday in each year.

33. The transport sector strike in February 1997 caused a 3,6% reduction on the level of production of this month, compensated in March and April. It does not show outliers above three standard deviations. The degree of uncertainty regarding production for the next month is 2,1%.

34. Likewise, useful information can be gained of the different industrial branches. It is easy to see that they show very different patterns of behaviour.

REFERENCES

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.

Chang, I., Tiao, G. C. and Chen, C. (1988), "Estimation of Time Series Parameters in the Presence of Outliers", Technometrics, 30, pp.193-204.

Hillmer, S.C., Bell W.R and Tiao, G.*C.* (1983), "Modelling. Considerations in the Seasonal Adjustment of Economic Time Series", Applied Time Series Analysis of Economic Data, U.S. Department of Commerce, Bureau of the Census.

Prothero, D.L. and Wallis, K.F. (1976),"Modelling Macroeconomic Time Series", Journal of the Royal Statistical Society, Series A, 139, Part 4, pp.468-85.

Revilla, P., Rey, P. and Espasa., A. (1990), "Automatic Univariate Modelling of Time Series: Application to the Industrial Production Indices", unpublished.

Wallis, K.F (1977), "Multiple Time Series Analysis and the Final Form of Econometric Models", Econometrica, v.45, n.6, September, pp.1481-98

Zellner, A. (1979), "Statistical Analysis of Econometric Models", Journal of the American Statistical Association v.74, n.367, September, pp.628-651.

Table 1

Series	Level Behaviour	Seasonal Behaviour	Working-Days Effect (%)	Easter Effect (%)	Feb 1997 Strike Effect (%)	Outliers		Uncertainty (%)
						Date	Type	
0	Trend	Yes	1.9	-4.0	-3.6			2.1
11	Local Oscillations	No	2.7					8.3
12	Trend	Yes				3/94 2/95 10/95 10/97 1/98	A. O. A. O. L. S. L. S. L. S.	13.0
13	Local Oscillations	Yes						6.4
14	Local Oscillations	Yes				9/93 11/93 5/96 8/93 3/98 10/98 11/98 12/99	L. S. A. O. A. O. A. O. A. O. A. O. A. O. A. O.	17.7
15	Trend	Yes		-5.4	-7.8			3.6
21	Local Oscillations	No				6/94 2/96 4/96 11/96 3/97 8/98 5/98	L. S. T. C. T. C. L. S. T. C. L. S. L. S.	11.4
22	Trend	Yes		-7.2	-6.7	1/94 12/95 8/96 9/99	A. O. A. O. A. O. A. O.	3.6
23	Trend	Yes	3.1					5.7
24	Trend	Yes	1.2	-2.2		2/93 8/93 1/94 12/94 9/96 3/97	A. O. A. O. A. O. A. O. T. C. L. S.	3.7
25	Trend	Yes	1.8	-4.1		12/93 10/95	L. S. A. O.	3.4
31	Trend	Yes	3.2			8/93	T. C.	4.5
32	Trend	Yes	2.9		-7.5	8/94	A. O.	6.2
33	Local Oscillations	No				5/93 7/93 1/94	A. O. A. O. A. O.	20.6
34	Trend	Yes	3.3			8/93	A. O.	5.1
35	Trend	Yes						13.0
36	Trend	Yes	3.5	-4.8	-12.2	8/93 2/94 4/94 8/94 8/95 8/96 4/98	A. O. L. S. A. O. A. O. A. O. A. O. A. O.	3.0
37	Local Oscillations	No				5/94 9/94 12/94 2/96 9/96 5/97 11/98 3/99	L. S. L. S. L. S. L. S. T. C. L. S. L. S. T. C.	6.5

Series	Level Behaviour	Seasonal Behaviour	Working-Days Effect (%)	Easter Effect (%)	Feb 1997 Strike Effect (%)	Outliers		Uncertainty (%)
						Date	Type	
38	Trend	Yes	2.5	-6.1		5/94 8/94 3/97	T. C. T. C. L. S.	7.0
39	Trend	Yes	3.0			8/94 6/99	A. O. L. S.	5.9
41	Trend	Yes	2.4		5.3	1/94 7/95 1/96	A. O. A. O. T. C.	2.7
43	Trend	Yes	3.2			9/93 8/99	L. S. A. O.	5.2
44	Trend	Yes	4.4			4/93 8/93 7/97 6/99	A. O. L. S. L. S. L. S.	5.9
45	Trend	Yes	3.1			4/93 2/94 9/94 12/95 4/96 9/97 5/99 6/99	T. C. L. S. L. S. T. C. A. O. A. O. T. C. A. O.	2.3
46	Trend	Yes	1.6	-5.3	-3.7	8/93 8/96	A. O. A. O.	3.9
47	Trend	Yes	2.1		-3.3			3.0
48	Trend	Yes	2.7	-3.2	-10.0	8/93 8/95 8/98 8/99	A. O. A. O. A. O. A. O.	3.7
49	Trend	Yes	4.2	-14.5		4/93 8/93 5/94 10/94 12/94 3/95 7/96 10/96 12/96 6/97 11/98	L. S. A. O. A. O. A.O. A. O. A. O. L. S. L. S. A. O. L. S. A. O.	5.4