

An Analytical and Graphical Approach for the Identification of Systematic Errors in Structural Business Surveys

*Un Approccio Misto, Analitico e Grafico, per l'Individuazione
degli Errori Sistematici nelle Indagini Strutturali sulle Imprese*

Angela C. Seeber, Tiziana Tuoto, Stefania Cardinaleschi

ISTAT
seeber@istat.it

Riassunto: Le procedure tradizionali per il controllo e la correzione dei dati non distinguono esplicitamente tra errori sistematici ed errori casuali. Tale distinzione si rende però necessaria sia per non ripetere errori evitabili nelle edizioni future della stessa indagine, sia per stabilire delle regole di correzione che non introducono distorsione nei dati. In questo lavoro viene proposto un approccio gerarchico per la costruzione del profilo degli errori nelle indagini strutturali sulle imprese, in cui si considerano livelli progressivi di esplorazione analitica e grafica dei dati, finalizzati all'individuazione di componenti sistematiche nei seguenti tre tipi di errore: errori di salto tra sezioni o all'interno di sezioni del questionario, errori che provocano incompatibilità aritmetiche ed errori che non danno luogo ad alcuna incompatibilità

Keywords: Error profile, Data quality, Explorative Data Analysis, Systematic Errors

1. Introduction

In structural business surveys, traditional data editing and imputation procedures evaluate the violations of single hard and soft edit rules, in order to identify all enterprises with suspect or impossible values. The identified large enterprises are reviewed interactively, while the remaining selected enterprises are corrected by a complex system of deterministic correction rules. However, in order to avoid the introduction of bias into the data, modern editing procedures distinguish between systematic errors to be treated in a deterministic way, extreme and influential random errors, to be reviewed interactively, and residual random noise, which can be treated by probabilistic error correction rules, at least in order to smooth data with respect to compatibility rules.

Following this modern approach, particular attention has to be paid on the construction of the error profile, in order to set up the corresponding correction rules as well as to evaluate the quality of data at the raw state and at successive stages of processing. In this context, the main goal of an efficient error profile is the identification of systematic error components in the data. By definition, systematic error is determined from a specific cause and introduces necessarily bias into the data. However, once the deterministic error generation mechanism is identified, systematic errors can not only be resolved by applying the inverse rule, but also avoided in future editions of the survey.

On the contrary, random errors are due to an infinite number of unknown causes or to an accidental mechanism and the success of their manual or automatic treatment can be considerably damaged if data is still affected by systematic errors.

In this work a general approach for the construction of error profiles in structural business surveys is proposed. This approach is based on the subdivision of errors into three types: skip errors between or within sections of variables characterised by the same measurement unit (number of employees, number of worked hours, monetary amounts), arithmetic fatal errors within sections and violations of soft edits, which have to be defined on the basis of the observed distributions of correlated variables (for example total worked hours and total labour cost).

2. The error profile construction method

The construction of error profiles involves necessarily the whole set of sampled enterprises, for which - at least implicitly – an error generation model can be assumed. In our context the target is to extract the deterministic parts of this model. The proposed error profile construction method is hierarchical, considering progressive levels of skip errors referred to sections, subsections and items of the questionnaire. In absence of skip errors in the involved variables, the presence of arithmetic fatal errors is investigated and box-plot distributions of their amplitude are analysed. Finally, characteristic relationships are explored by simple scatter plots on those enterprises with no fatal errors in the involved variables.

This approach corresponds to subdividing the set of questionnaires according to a tree structure, whose final nodes contain specific error situations which can be analysed with respect to systematic components. In particular, the first ramifications follow conditions related to the presence of not empty sections, subsections and items of the questionnaire. At the successive levels the violations of arithmetic hard and soft edits were investigated on those enterprises which satisfy the conditions of their parent node. The whole procedure for the construction of the profile is illustrated in table 1.

Table 1. *Hierarchical levels of the error profile construction method*

Condition	Hierarchical Level	Systematic errors investigated
Questionnaire not empty	Compilation of questionnaire sections	Particularly high frequencies of skip errors between questionnaire sections
Sections not empty	Compilation of subsections or single items	Particularly high frequencies of skip errors between subsections or single items
Items not affected by skip errors	Satisfaction of arithmetic rules involving subsections or single items	Particularly high frequencies of balance errors in both directions or in one particular direction
Items not affected by skip errors	Box-plot distributions of balance error amounts	Asymmetric distributions or distributions not centred at zero
Items not affected by any fatal errors	Scatter plots on pairs of correlated variables	Clusters or anomalies in the scatter plots

3. Data and results of the application

In this work we tested the method on the raw data collected in the Labour Cost Survey 1997, consisting in a sample of 11.380 enterprises. The survey is performed every three years and its aim is to collect harmonised information for the European Union labour market. The target population are all enterprises on the Italian territory with at least 10 employees operating in the industry or service sector.

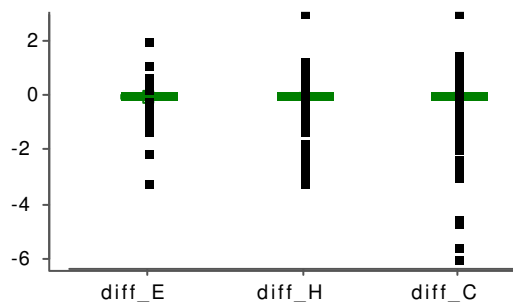
The present application involved the two main sections of the questionnaire, the occupation structure section and the labour cost section. The first section contains, for each professional category and as total amounts, the monthly mean number of employees, the mean number of part-time employees and the annual number of hours worked. The second section collects the total amount of labour cost, subdivided with respect to the main components of salaries and social contributions.

Table 2: *Number of enterprises according to the error profiles referred to main totals*

	Employees	Part Time	Hours	Cost
Section not empty	11367	11367	11367	11352
→ Only total compiled	4	7	37	2
→ Only partial items compiled	756	613	983	799
→ No skip errors in partial items	623	486	769	697
→ Total and partial items compiled	10595	4898	10188	10550
→ No skip errors in partial items	8821	4021	8675	10468
→ Total = sum of items	8463	3991	7829	7730
→ Total < sum of items	131	9	520	1660
→ Total > sum of items	227	21	326	1078

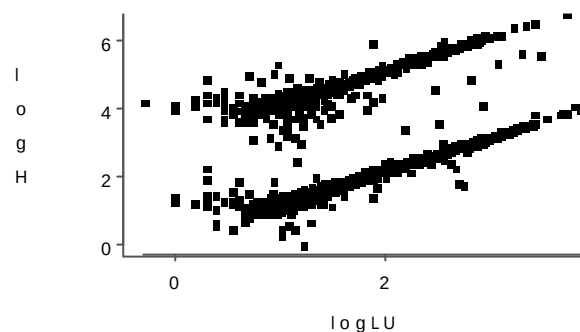
Table 2 shows the part of the error profile referred to the three main totals of the first section, number of employees, number of part-time employees, number of hours worked and to the main variable of the second section, total labour cost. As for skip errors between totals and corresponding partial items, in both sections there is a particular high amount of respondents which compiled partial items only. In the employment section, there is furthermore a high incidence of skip errors between partial items. The complete error profile shows that these errors are mainly due to zero values in items referring to hours worked, evidencing the difficulty for enterprises in retrieving the necessary information. As for arithmetic errors, no strong evidence for systematic violations of balance edits can be observed.

Figure 1: *Box-plots of the amounts of balance edit violations*



This is confirmed by figure 1, showing the box-plot distributions of the amounts of the observed balance errors, evaluated as the difference between the logarithmic transformations of the totals and the corresponding sum of partial items. However, evident systematic measurement errors can be observed in the variables referring to hours worked and labour cost, due to a misunderstanding of the requested measurement unit. Figure 2 reports the logarithmic scatter plot of total number of worked hours against total number of labour units (total number of employees – half the number of part-time employees), evidencing a distinct cluster of enterprises failing the division by thousand indicated by the questionnaire compilation rules.

Figure 2: Scatter-plot of Hours Worked against Labour Units



4. Conclusions

Our application showed the importance of reporting fatal errors and identifying soft errors in a hierarchical manner, in order to distinguish between skip and measurement errors as well as to investigate hard and soft measurement errors in a graphical way. The strength of graphical data exploration lies furthermore in the opportunity of identifying simultaneously systematic errors and extreme random errors or outlying values to be revised interactively by experts. The proposed error profile construction method can be used for repeated applications in order to check the hypothetical causes of the identified systematic edit violations and, more in general, to monitor successive actions as the editing and imputation process goes on.

References

- Granquist, L. (1995) Improving the traditional editing process, in: *Business Survey Methods*, Wiley N.Y.
- Houston, G., Bruce, A.G. (1993) Interactive Graphical Editing for Business Surveys, *Journal of official Statistics*, vol.9, nr.1, 81-90
- Cirianni A., Di Zio M., Luzi O., Seeber A.C. (2000) The new integrated data editing procedure for the Italian Labour Cost survey: measuring the effects on data of combined techniques. *Proceedings of the International Conference on Establishment Surveys II*, Buffalo, June 17-21 2000.