Outliers Identification and Reliability Evaluation of Provisional Business Survey Data for National Accounts

Vincenzo Caponi
ISTAT, National Accounts
vincenzo@caponi.net

Augusto Puggioni
ISTAT, National Accounts
Via Depretis 74/B
Rome, Italy
puggioni@istat.it

1. Introduction

Business surveys constitute one of the main sources used by ISTAT for National Accounts (NA) estimates. A database is in fact prepared from these surveys in line with NA definitions and concepts and consistent with the methods used for compiling NA. This means that the information is first of all analyzed by studying the quality (and consistency) from the NA angle, so that the data from the survey is "fitted" to NA schemes and needs, with the objective of maximizing its quality (Calzaroni, Puggioni, 2001). This can involve data editing and imputation methods different from those used for surveys since the aims could be different. One of the most important surveys for non definitive estimates of GNP is the *Provisional Estimate of the Value Added of Enterprises* (PEVAE). The survey collects information on enterprises with at least 100 employees operating in the industrial, trade and service sectors. The main variables, collected according to the enterprise's main activity, are: costs and revenues, employment and investments. The data regarding year t starts to be available from the month of February of the year t+1 and completed by September. In 1999, the survey covered 55% of a universe of around 8,000 enterprises. More detailed data, also available for secondary economic activities, is instead collected by the *Enterprise Accounts System* (EAS), the results of which are however available 18 months after the target year. The information provided by the PEVAE survey regards both the year in question and the previous one. The main problems in using the survey for NA estimates are the following: 1. Lack of information on any of the enterprise's secondary economic activities. 2. Unavailability of some items included in the calculation of aggregates according to SEC 95 definitions. 3. Presence, in the survey referred to year t, of data still not definitive for the same year t. The first problem is resolved with a good approximation by applying the last structure by the EAS survey to the PEVAE data. The second is actually irrelevant, since we are interested in estimating the annual variations of the per capita values, having checked that the variables calculable by PEVAE are under this aspect good proxies of those requested by SEC. For the third problem, it was decided to use an econometric method to reduce the error caused by using non definitive data, given the good

number of cross-sections data in the survey. The method described in the following paragraph also enables the identification of outliers, in the sense explained below.

2. The Model

The PEVAE survey has a part with questions on the target year, which for simplicity we call provisional data, and another regarding the previous year, which we call definitive. The analysis we propose here is based on the consideration that the definitive data is generally more reliable than provisional data. We define *estimate error* the discrepancy between the provisional (i.e. that estimated by the enterprise) and the definitive data available the following year. Our goal is thus to find a procedure which permits us to define the probability that enterprises commit these errors, their extent and, finally, that gives us the possibility of correcting them. Since we possess the data of all the enterprises we can compare the data of each one with the rest of their population, then we create a new variable representing the deviation between the effective value added of each enterprise and the average of the branch to which belongs. Thus, following the microeconomic law which says that in the long run similar (those belonging to the same branch) enterprises should have similar productivity (i.e. value added), we assume that those firms with a larger value of this new created variable with more probability provided erroneous provisional data, and that greater is the value greater is the error itself. Then, for the year previous to the one to which we wanted to correct, we did a regression analysis between the final and provisional value of the per capita value added. Through this regression we can thus have an idea of how the final data has actually changed compared to the provisional data and in what direction. The regression model also takes into account the heteroscedasticity of the error that is, expecting a greater error the greater the distance of the enterprise's value from the branch average, we also expect that the relative variance from the same is correlated to this latter variable. Hence we use a model that treats a non *white noise* error. Finally having obtained the model, its parameters are calculated and used for correcting all the provisional observations. Once again the correction takes into account the deviation from the average of the branch variable. On this basis a series of probabilities is constructed as a positive, specially built, function of the extent of the deviation. The correction thus depends on the probability, that is the greater the probability that the provisional observations are erroneous, the greater they are modified.

REFERENCES

Davidson, R., McKinnon, J., (1993). *Estimation and Inference in Econometrics. Oxford University Press*

Calzaroni, M., Puggioni, A., (2001). Evaluation and analysis of the quality of the National Accounts aggregates. *Metodi e Norme, ISTAT*

RESUME

Ce document présente une méthode économétrique qui permit d'utiliser correctement les bilans provisoires des entreprises pour les premières évaluations des comptes nationaux.