Società italiana di economia pubblica
XV CONFERENZA

DIRITTI, REGOLE, MERCATO
Economia pubblica ed analisi economica del diritto

Pavia, Università, 3 - 4 ottobre 2003

INFORMATION SOURCES AND METHODOLOGICAL ISSUES IN THE DIECOFIS PROJECT

M. GRAZIA CALZA, FRANCESCA INGLESE

EUROPEAN COMMISSION

INFORMATION SOCIETY TECHNOLOGIES
(IST) PROGRAMME

DEVELOPMENT OF A SYTEM OF INDICATORS ON COMPETITIVENESS AND FISCAL IMPACT ON ENTERPRISES PERFORMANCE

INFORMATION SOURCES AND METHODOLOGICAL ISSUES IN THE DIECOFIS PROJECT[1]

Maria Grazia Calza and Francesca Inglese

ISTAT

[1] This work draws on research carried out under the EU FP5 Project DIECOFIS (IST-2000-31125) which aims at developing a system of indicators on competitiveness and fiscal impact on enterprise performance. The project has four components focused on the development of 1) a comprehensive, high quality and robust, multi-source, integrated and systematised data base, which can be relied upon for policy impact analysis and to estimater a broad range of micro-founded statistical indicators. 2) first generation microsimulation models that can serve to monitor and simulate the impact of public policy on enterprises; 3) a system of micro founded indicators on competitiveness that can be used to map and benchmark enterprise performance; and 4) prototype national corporate tax Microsimulation models and move towards the development of an EU-demonstrator.

XV Riunione scientifica Siep
3-4 ottobre 2003

INFORMATION SOURCES AND METHODOLOGICAL ISSUES IN THE DIECOFIS PROJECT

(preliminary draft)

M. Grazia Calza,
Istat, via Magenta 2, Roma, e–mail: calza@istat.it

Francesca Inglese
Istat, via Magenta 2, Roma, e–mail: fringles@istat.it

Summary

During the last quarter of a century the increasing availability of micro data on individuals and families and the development of longitudinal and cross-sectional micro-simulation models has brought tremendous progress in the analysis of the impact of public policy on households. Enterprise micro data and micro-simulation models for the analysis of the impact of public policy on business have not known a similar advance. Most often this is due to lack of micro business data because they are either non available or because they cannot be accessed for confidentiality reasons. On the whole, the field of enterprises micro-simulation models remain a much under-researched and underdeveloped area of study in EU countries. Against this background policy requirements seems to necessitate different sorts of models operating on quite different sets of data and for some purposes micro-data from different sources need to be linked. In other words, a range of databases and methods may be used but that for many requirements assembling a suitable micro-database will involve the record linking of data from various sources, including tax administration, statistical surveys, and accounts.

In this field of analysis DIECOFIS project aims at making an important contribution in the development of micro-simulation prototypes for the business sector, both in a national and EU perspective, that can be innovative for their potential in showing the impact that policy have on competitiveness and economic renewal Access to high quality enterprise micro data information systems will prove crucial. This involve overcoming complex problems, such as restricted access, privacy guarantees, methodological issues and development of IT tools.

In this paper we describe the main issues in the creation of a multi-source integrated and systematised data base of enterprises data. We analyse the main statistical objectives of the integration of different sources of data and the methodologies to implement it; the fiscal background for micro-simulation models; the preliminary stage of the integration process; the analysis of the problems generated by combining administrative data and surveys data to prepare the matrices for the micro-simulation purposes and the proposed methodological solutions.

1. Introduction

The integration of different sources of data, both statistical and administrative, has been the object of much research. This exercise can satisfy many objectives of both a statistical and a non-statistical nature. Among the main statistical and economic objectives of data integration from different sources we can find: *the construction of archives, the improvement of data quality, the development of economic analysis.* For instance, the construction of the ISTAT Archive of Active Firms (ASIA), aimed at creating a list of statistical units to be used as a sample basis or as a reference population for censuses, is performed by uniting in one single list the enterprises listed in different administrative sources. Information from administrative source is a precious resource to be united with statistical information obtained with direct methods, i.e. through surveys; because it can contribute to the improvement of data quality, making the control on cover and answer mistakes more effective in surveys and censuses if combined with the usual techniques of allocation and editing. Eventually also the development of economic analysis at micro and macro level (micro-simulation, impact analysis of policies, systemic analysis and indicator construction) can be carried out by joining two or more different sources in order to obtain one single database with more information.

Generally the integration process of various sources for the development of integrated systems can be performed through three different methodologies: merging, record linkage and statistical matching. The first and the second deal with the identification of the units in two or more different files, the third deals with the problem of integration when units in different files are not the same. When integration deals with sources whose units are identified through an univocal identification code matching can be done with a simple merging; if the involved units do not have an univocal identity code, connection criteria among different information coming from different sources have to be established: in this case record linkage techniques can be used (exact matching). These techniques have the scope of identifying couples of records from two data sources related with the same unit (Jabine and Sheuren, 1986). If the sources do not have the same unit, but they have common information, then statistical matching can be used in order to collect information of similar units respecting some criteria (Pass, 1984).

The construction of integrated databases is strongly conditioned by the availability and quality of information, as well as by the sample and non-sample nature of the involved data sources. It can also be performed at a micro level if we refer to single units, or to the macro level, if we refer to macro-aggregates. Even the choice of integration methodologies to be used depends on the above mentioned factors. The objective of the construction of integrated data-sets for economic analysis is to combine information that are available in one source and lacking in others.

Regarding the DIECOFIS integrated data-base the logic and the choices behind the structure and the adopted methodological solutions for the integration procedures if, on one hand, try to answer to the information requirements of the micro-simulation models for the analysis of the impact of public policies, on the other, allow to develop a microsimulation approach that is users' need and demand driven. Section 2 gives a brief description of the microsimulation modules and discuss their information requirements.

The reminder of the paper is structured as follows. Section 3 describes the sources utilised in the construction of integrated data-sets; section 4 investigates the preliminary phase of harmonisation of the integration process; then, in the section 5 we describe the problems determined by combination of data from the different sources and we analyse

the further step of the integration process which deals with the following issues: the statistical analysis, the data processing, the estimate of population parameters.
Finally, we give a brief description of how the sensitivity analysis may be used to validate our integrated system.

2. Main information requirements of the micro-simulation models

The structure of the database and some methodological solutions for the integration procedures were adopted to fit the information requirements of the micro-simulation models. Regarding fiscal micro-simulation, the general aim is to evaluate the effects of the Italian tax system on enterprises' decisions developing different modules of analysis dealing both with indirect taxes (VAT and taxes on production) and with taxes on revenues created by the enterprise (Irpeg, Irap) as well as social contributions, that are still an important part of labour costs. The simulation techniques for the different aspects of the fiscal provisions are subjects of specific deliverables of DIECOFIS project. In this part of the work we just analyse problems arising from available data and the possible limits they impose on the analysis. Regarding simulation of taxes on revenues created by enterprises (Irpeg and Irap), in the perspective of constructing the database, the target is to reconstruct the tax base of the corporation income tax (Irpeg) and of the regional tax on productive activities (Irap). Regarding the simulation of social contributions, the model refers only to contributions due from private employers for regular employees. In all cases we deal with static models that do not consider changes in the behaviour of the involved economic subjects. This means that we evaluate only the first impact of fiscal policy leaving behind the further effects of adjustment. Nevertheless, the database wideness and the typology of the models make the structure quite flexible, giving a chance to simulate various scenarios and to adapt the model to different tax rules. On the basis of these first models we might think about dynamic and behavioural developments.
The microsimulation of the impact of fiscal policies on firms deals with the following modules: Irap, Irpeg and Social Contributions

1) Irap is a regional tax on productive activities in force in the Italian tax-system since 1998. It is a low rate tax with a wide tax base dealing with the net value added (or product). They are passive subjects of the tax all subjects that carry out an independently organised activity, directly aimed at the production or the trading of goods or the provision of services. Ira pos paid by all the enterprises (individual, person, corporate, including non-commercial entities), art dealers and public administrations. The tax base is defined in a different way for different passive subjects on the basis of their economic sector, in order to approximate as much as possible the value added of the various taxpayers. For societies subject to ordinary accounting rules other than financial, credit and insurance concerns, the taxable value is established in a detailed way referring to the items of the economic profit as provided by the balance sheet. The value added is determined by subtracting from the production value (revenues, variation of leftovers and works in progress) some production costs such as the purchase of gross materials, goods, amortisation and material immobilisation. Personnel expenses and interest payable are non deductible. Apprenticeship contracts expenses, 70% of the cost of training contracts and expenses related to disabled workers are exceptions. For enterprises with their residence in Italy the value of production realised abroad is not taxable. Regions have the power of changing the rate, increasing or diminishing it up to a maximum of one percentage point as well as differentiating it according to activity sectors and passive subjects categories. Therefore in order to establish the taxable base

for Irap, we need data relating to the value of production and the costs of production. This detailed information is provided by the data coming from the ISTAT survey on large enterprises (LE) and from the survey on small and medium enterprises (SME). Analysis can be performed on a data set which includes all the surveyed statistical units (about 56 thousands units of analysis in 1998, all large enterprise and a sample of SMEs), with the addition of some auxiliary information acquired from the ASIA archive through merging techniques, and variable harmonisation. Using the weight attributed to each unit of the SME sample we can equalise the number of enterprises estimated from the sample to the one in the universe. Finally, for 1998 simulation the number of Irap taxpayers considered in the model is around 4 million.
In order to evaluate public finance manoeuvres we should consider that some of the items contained in enterprises balance sheets and then reported in the questionnaires of the surveys do not coincide with fiscal provisions. Therefore, the main problem is the reconciliation of balance-sheet and fiscal values in order to reproduce in the model the changes that the Tax Authority requires for some items recorded in the balance-sheet. That is why it is necessary to compare survey data relating to certain variables with a sample of the corresponding fiscal data, provided by the Tax Autorithy. On 1998 enterprise data the Irap rate is 4.25% and it affects all enterprises of all considered sectors in all regions. That is, no dimensional or sectorial allowances could be used in this year.

2) In Italy taxation on enterprise revenues, i.e. net revenues in the taxable period, is different according to the juridical nature of the enterprise itself: individual enterprises and companies of persons on one side, and corporate enterprises on the other. Enormous difficulties linked to the possibility of inferring the taxable basis for companies of persons, circumscribed the study on taxation of enterprise revenues only to corporations. They are passive subjects of Irpeg Tax (Corporate tax) limited companies, limited partnership with share capital, limited liability companies, mutual aid society and co-operatives, public and private bodies different from companies, both having or not as exclusive or main object the performance of trade activities, companies and bodies of any kind with or without legal status with no residence in the state territory.
The taxable income is reconstructed in an analytic way by algebraic comparison of the positive and negative components of the profit and loss account, distributed according to a principle of correspondence with the fiscal period. From the taxable income we subtract deductible expenses, apply the taxable rates and we obtain the gross tax. From the latter, using detractions, the result is the net tax, eventually subtracting tax credits we have the tax to be paid. If the calculation of taxable value results to be a loss, the latter can be deducted from revenues of posterior fiscal periods, not beyond the fifth year.
For the reference year of Irpeg simulation, 1998-1999, the tax rate is 37%; the 2001 Budget Law lowered the IRPEG rate to 36% starting in 2001 and to 35% starting in 2003. A complication in calculating taxes on the reference year of the simulation (1998), is related to the introduction since 1997 of a dual taxation regime for enterprises. They are under a tax regime named Dual Income Tax (DIT) introduced with the purpose of reducing both the discrimination against equity finance and the effective tax rate. This modifies the method to determine income taxes of individual and juridical subjects (Irpef and Irpeg) introducing favourable treatment on a portion of this revenues. More in detail, facilitated treatment is based on the division of taxable income in two components: “normal” income, attributable to new capital formation, taxed with a reduced rate of 19% and the “residual” income taxed at the ordinary rate. The portion of income or profits enjoying a reduced rate, corresponds for corporates to the value of additional stock issuance evaluated in consideration of the value the stock had when the

fiscal year was closed (30 September 1996). Such income is the result of the net increment of stock invested by the company compared to the existing stock in September 1996, a figurative internal return rate is provided by the law. Therefore, for the calculations, the database must contain longitudinal information. In more detail, for corporate companies the cash given by the partners, including deposits without security, allocations to fill previous losses and undistributed profits must be collected from 1996.
Furthermore, the construction of the Irpeg taxable basis requires that the reference data-set includes for some variables relative values of the previous year. It deals especially with variables used to calculate amortisation of material immobilisation.
For the purposes of our analysis the diversity between civil and fiscal criteria used to determine taxable income in a fiscal period, is a limit of the data available both from surveys and from Chambers of Commerce. The base to determine enterprise profits is the net profit (or loss) resulting from the profit and loss account, which is a variation of addition or subtraction reflecting the divergences between civil and fiscal doctrine. Unfortunately information contained in balance sheets drafted for civil purposes do not allow to consider the majority of variations in addition or subtraction. Thus, in order to calculate fiscal adjustments in the micro-simulation model it was necessary to transform the accounting values into fiscal values on the basis of the aggregate corporate tax returns published by the Tax Authority for the fiscal year 1998. Estimates on year 1998 have been performed on a data set which includes all large corporate companies for which the information from the survey data have been integrated with some information from the administrative source the Chamber of Commerce. New estimates from year 1999 and onwards will be performed on a data set which should includes also small and medium corporate companies. Many statistical problems arise from the construction of the corporate database, the survey data on small and medium enterprises do not report information on assets and liabilities so in order to estimate the corporate tax we need to integrate this data with the administrative data on balance sheets and thus we need to reconcile variables from profit and loss. There are also problems of imputation of missing data[2].
In the perspective of expanding data sources and create the overall database for policy impact analysis, the integrated system will acquire from the fiscal source, for the units of interest, information on the variables that constitute the link between "civil balance" and "fiscal balance" in order to have the true basis of enterprise taxable income. To these link variables we should than add other variables, mainly, fiscal ones (such as the incidence of facilitation, or the use of the benefits of the Dual Income Tax, etc.) in order to better calibrate the model in terms of behavioural choices and of total analysed tax revenues.

3) The analysis on the simulation of social contributions paid by employers for dependent workers requires as first step the construction of the tax base for social contributions. This calculation can be obtained by classifying the employers in the following categories: executives, white collar workers, manual workers and apprentices and to have information on the number and on wage and salaries. Given the tax base for social contributions the contribution rates are applied in order to compute the employers' social contributions. Specific rates are applied for sector of activity, firm size, type of contract and contribution. About the information requirements, the construction of this module requires the acquisition of a new administrative source, that is the archive of the Nation Institute for Social Security. In fact, the tax base cannot be calculated using only the information provided by the surveys sources and the commercial accounts data.

[2] For more details on this see Denk, M. and Oropallo, F. (2002).

3. Description and analysis of the sources

The sources used to construct the integrated data-sets are: the register of active businesses (ASIA), the two statistical surveys on business profit and loss account and the administrative data on commercial accounts by corporations company and enterprise income tax-return. ASIA, constructed by the integration of administrative sources represents the universe of active businesses (about 4 millions of enterprises); information related to the units of the archive are of an identificational kind (name, fiscal code number, legal status), of a structural type (main and secondary activity sector of the enterprise as expressed by ATECO91 code with 5 figures, workers, location, etc.) of a demographic kind (date of activity begin, retirement, etc..). The survey on enterprises with more than 100 workers (LE) is an exhaustive one and provides a complete picture of the economic results of the Italian enterprises' system on industry and services excluding the financial brokerage sector. The survey on small and medium size enterprises (SME) is based on samples, the reference universe is made of enterprises with 1-99 workers and the observed field includes all economic activities such as industrial, commercial and services (for further details see the studied case). Statistical information collected in the two surveys enables to identify the Value Added as well as other aggregates that are necessary to estimate National profits and loss accounts as well as the inter-sector table of the Italian economy.
The statistical information collected in the two surveys is wider and more detailed in the former than in the latter. As far as it concerns the small and medium enterprises information related with the statement of assets and liabilities is less complete.
Moreover, while for the former a reconstruction is possible, at the statistical units level of longitudinal information, for the latter, since there is not a panel sample, it is not possible. Two observed aspects, partially due to specific objectives of the two surveys and to the sample used for the SME surveys, explain the need to use added sources to estimate the econometric model of micro-simulation (Irpeg).

	LE survey		SME survey	
Year	1998	1999	1998	1999
Corporate enterprises	7,124	7,339	15,372	15,329
Non corporate enterprises	1,330	1,395	32,112	30,618
Total	8,454	8,734	47,484	45,947

Administrative data related to commercial accounts, lodged at the Chambers of Commerce, provide information on the profit and loss account and the statement of assets and liabilities of corporate companies, in some cases with more detail than in the surveys.
Information relating to the administrative archive of the Ministry of Finance, i.e. enterprise income tax return year 1999 are not included in the integration process, but are going to be used in order to verify the quality of estimated Irpeg and Irap in the micro-simulation process and to calculate fiscal adjustments. Fiscal data availability is complete for enterprises with 100 and more workers while is limited only to some economic activity sectors in the case of small and medium enterprises, therefore only some aggregates will be verified.
In the perspective of expanding data sources, it would be useful to acquire also data from The National Institute for Social Security (INPS) on the statistical archive on the Osservatorio sulle Imprese. The integration of this administrative source is necessary to develop the social contribution module as some computations cannot be performed using the survey data and commercial accounts data.

4. The preliminary stage of harmonisation

The objective of the preliminary stage of integration is to perform the data harmonisation acquired from different sources. Harmonisation deals with the definition of statistical units, of the interested population and of the variables. Regarding units harmonisation, the records referring to the same definition of unit have to be selected; for the harmonisation of the population we need to select those records referring to the same population of interest; for variable harmonisation, common variables have to be established in the same way.

The ASIA archive was the hub of the integration process in the preliminary stage. The archive allowed to harmonise units from different sources and to harmonise the interested population. Through a linkage operation the observed units in surveys on enterprises were found in the ASIA archive using as coupling key the "fiscal code" number and the VAT registration number. For both year 1998 and 1999 the linkage work was carried on through a simple merging operation of the surveyed units through surveys of those years. Since merging of units did not happen for all the units, for some of them the linkage was made through the ASIA archive of the previous year. This solution is due to the fact that the survey of the reference population for one year is made on the active businesses of the previous year. Auxiliary information were added to the mentioned enterprises on location, economic activity, legal status and total number of workers, some of them were present in the archive but not in the surveys.

Regarding the data-set construction for the Irpeg module, since the analysed statistical units are made by of corporate companies, harmonisation of the interested population was necessary. In order to do that a sub-population of classified enterprises was selected from the ASIA Archive, on the basis of their legal status.

In order to acquire commercial accounts data we had to refer to the population of corporate companies. In order to have information on the units of the two sources (surveys and commercial accounts), all the corporate companies present in the LE and SME surveys were singled out. In a further operation the so identified companies were excluded from the population of corporate companies and from the remaining ones a stratified sample was selected, according to the ateco sector , the number of workers class, and the administrative region. In this way a list of enterprises was built including the corporate companies present in both surveys as well as those selected by the sample. Eventually, through their identification code, the units selected from the commercial accounts archive of 1999 were extracted. The same list was used for the extraction of units from 1998 commercial accounts in order to obtain longitudinal information on the units themselves.

An important operation was the harmonisation of variables observed in the two surveys and afterwards among the variable of the surveys with those common in the commercial accounts. Surveys, as mentioned, present some diversities with regards to the acquired information, above all with reference to the detail that compose macro-variables and the absence of some information in the SME survey. Regarding the profit and loss account section important differences were not recorded in both surveys for they only lack some variables such as profits from shareholdings, financial profits and expenses, revaluation, devaluation, etc. and of their corresponding details in the SME survey. The section dealing with the assets and liabilities statement is not very detailed in the SME survey, its variables are: intangible assets, tangible assets, long-term investments, available cash, risk fund, the TFR and some other items of the leftovers. Regarding employment and personnel expenses in both surveys macro-variables are present but some details are missing in the LE survey, and some others in the SME survey. In the section dealing with acquisitions and assets in the fiscal year the LE survey is more detailed also in the

section “other information” (investments, credits and debts) and in the section recording information on employment located in the local units in enterprises with various locations.
In some cases, in order to adapt the variables we had to aggregate some items. The harmonisation of survey information with commercial accounts information was performed on the common variables. In some cases, in order to harmonise the information we had to aggregate some voices reported in the surveys. From a general point of view information in the archive are less detailed than those collected in the surveys regarding the profit and loss account, but are much more detailed regarding the loss and liabilities statement especially as far as it concerns the SME survey. Other information are completely missed.

5. Combining data based on different sources

5.1. Preface

After the above-described stages have been completed, the situation determined by the combination of data from the different sources has to be investigated. In general, matching of more sources at the record level can be realised partially or not at all. The lack of record matching, from a methodological point of view, can assimilated to the statistical problem of missing data. According to different situations, it can be treated as a typical problem of allocation of missing data or as a problem of data-fusion where the missing observation have a special structure (for instance two samples with different units, some common information and others present in both).
The output created for the Irap and Social Contribution modules was obtained uniting the two files containing the surveys after defining and harmonising the variables of the micro-simulation model. While the creation of the output for the Irpeg module required the integration of statistical and administrative sources. The fusion of the two sources with merging technique at the record level, realises a partial matching. So it is possible to acquire some information, lacking in the surveys, from the commercial accounts for those units that have been matched. We distinguish two different situations to analyse and to check the data. For the couples of records in common between the two data sources, the data check objective is to identify the systematic errors and the differences on the common variables collected in both sources. For the non-matched units occurs to analyse a different type of error generated by missing data (5% for LE survey and 20% for SME survey). The analysis and the treatment of the two statistical surveys matched by commercial accounts data are conducted in a separated way.
The final objective of the integration is a new data-set with more information and a better data quality. In fact, the treatment of the discrepancy on the common variables from different sources allows to choose those values that better represent the phenomenon analysed and to reduce non sampling errors. The reconstruction of the missing value in SME survey contributes to reduce the sampling error.
Once the complete matrix for the Irpeg module has been reconstructed, both for the couples of records in common between the two data sources and for those records in which missing information has been imputed, consistency at the micro level has to be verified, that is, that the editing rules, the rules linking two or more variables, have not been broken. Furthermore the analysis on the distribution of the interested variables has to be conducted to identify possible statistic errors.

5.2. Statistical analysis and data processing

The aim of data analysis on the acquired sources was above all to check information consistency on some main variables of a macro-type, i.e. harmony among total values (of the macro-type values) with the total, calculated aggregating the value of analytic voices.
This type of check was performed on the basis of the main variables as reported in the following sections of the questionnaire:

- In the section on the profit and loss account we checked the revenues, leftovers variation, production value (purchases, services, personnel, amortisation, etc.);
- In the section on the statement of assets and liabilities we checked intangible and tangible assets, credits and debts.
- In other sections we checked variables such as total employment, personnel expenses, acquisition of assets in the fiscal year, etc.

The same check was performed on commercial accounts regarding some of the main variables of the profit and loss account and of the statement of assets and liabilities.
Another important check was performed on the total value of some variables at an aggregated level according to the number of workers class, the ateco categorisation and the localisation to verify differences between values calculated on the same units present in the surveys and the administrative archive and to verify the presence of systematic errors.
The alignment of information contained in the archive and of those collected through in the surveys is verified at an micro level to evaluate the strategy to reconstruct the integrated data-set. The comparison among the commercial accounts data and the survey data was performed on the main items of the profit and loss account, in a separated way for enterprises with more than 100 workers of the LE survey and for small and medium size enterprises of the SME survey, with no missing values in the commercial accounts items. The used variables were: production value, production costs, financial profits and expenses, value adjustments, extraordinary profits and expenses, taxes on profits and after-tax profits and others common variables. The comparison was performed creating classes on the basis of the size of the percentage gap CA/LE (or SME) – 1. In most cases a good alignment of information between the two sources was recorded. The highest portion of enterprises is contained in the variation class (-2, +2). When the discrepancies are relevant it occurs, on the base of some criteria, to select one of the two values.

5.3. Imputation techniques: parametric and non parametric methods

Regarding the reconstruction of the missing data, as the mechanism that generates missing data is ignorable, that is it is of the MAR type (missing at random), we can apply standard analyses for incomplete data (R.J.A.Little and D.B.Rubin -1987) this is because the two sub-populations, that is the one with exhaustive data and the one with missing data, are not characterised by different distributions of the variables with missing data. Through the use of covariates it is possible to identify the characteristics of the units with missing data so that missing information can be recovered by considering the units with complete data.
Probabilistic imputation methods can be distinguished in parametric and non parametric: the former are defined by a model, the latter by a donor. In the first case the value of the missing datum is estimated by applying a model chosen and formulated on the basis of the covariates and hypothesing a probabilistic relationship between the variables with missing data and the covariates. The model permits to estimate the

probabilities with which every value of the variables with missing data occurs. Therefore the imputation of possible values occurs in casual fashion within the same class of covariates.
Through the second method imputation is performed taking the information from a unit with complete data that is similar to the unit with missing data. It is possible to verify that for the same record to be corrected more than one donor is available so that it is necessary to randomly choose one from within the available group.
The imputation through a model guarantees respect for the distribution of the simple variables and of the joint distributions, at least with regards to the covariates used in the model. This method requires careful analysis of the phenomenon for the choice of the most efficient model. The imputation performed using the donor methodology, based on the principle of similarity of behaviour, consists in the search of a donor (similar unit) that is performed through the analysis of the covariates that makes the search for a donor more efficient and effective.
This technique guarantees a smaller distortion of the mean, it does not produce distortions either in the simple distributions or in the conjoined distribution, and finally, it does not flatten replies around the mean, safeguarding the variability of phenomenons. The similarity between donor units and receiving units is found on some selected matching variables on the basis of their correlation with the variables to be imputed. The concept of similarity is translated in mathematical terms as a distance function. Imputation based on the donor method with minimum mixed distance (chosen for the solution of our problem) aims at identifying, for every variable with missing data, a donor unit by calculating a mixed distance function on a restricted set of units, eliminating immediately those with greater elementary distances.
The weighting of the distances is carried out on the basis of the weights taken from the chi-square tests of independence.
The variables used to establish similarity between units may be distinguished in: strata variables and matching variables. The former allow the identification of groups of units that define sets that are not similar. The latter allow to identify inside one and the same stratum the donor unit closest to the receiving unit.
Two typologies of imputation techniques can further be identified with regards to the criteria for the reconstruction of missing information: simple imputation and multiple imputation. The simple imputation techniques consist of the substitution of one single value for every missing value. Multiple imputation techniques consist in the substitution of every missing value by more acceptable values that represent the possibilities distribution (Rubin 1987). The advantage of single imputation is that it utilises standard analytical methods on complete data, even on data sets with imputed data. However, it may lead to an erroneous estimate of the sampling variability as it does not consider the variability of missing values and the uncertainty deriving from the lack of knowledge of the most appropriate non response model. Multiple imputation consents, when imputation represents repeated casual extractions under a non response model, to obtain valid inferences that consider the additional variability due to missing values only through the combination of inferences on the complete data. Multiple imputation methods for multivariate data presuppose a missing data mechanism that can only depend from observed values of MAR (Missing At Random) type. Repeated imputations are obtained through a system of simultaneous regression models, based on one single multinormal model, in which, every variable is potentially dependent on all others. The simultaneity of estimates is made possible by the use of the iterative algorithms: EM and Data Augmentation (DA) (Schafer, 1997) while the first one provides preliminary estimates of great accuracy, the second that is a particular type of Monte Carlo Markov Chain (MCMC) converges towards the predictive Bayesian distribution after which multiple imputations are generated whose convergence speed

depends on the rate of missing information. The multiple imputation for quantitative data can be performed through two methods: a parametric method (Predictive model based method) that performs Bayesian conditional imputations on normal regression models, and a non parametric method (Propensity Score Method) that performs conditional imputations with the donor method, with imputation classes formed on the basis of the level of propensity to non response determined through a logistic regression model. While the parametric model may be used for every missing answer structure, the non parametric method may be applied only to values missing in monotonous fashion (that is in the case in which in a unit a missing value is followed only by other missing values).

5.4. Calibration methods to estimate the population parameters

For the estimation of the variables of the Irpeg model we will need to employ weight adjustment techniques. In fact the units of the matrix are a subset of a sample (SME) in which weights have been constructed on the basis of the sampling and estimation strategy of the survey that does not consider in the stratification the legal form of the units and as a consequence does not take it into account in the system of weights to be associated with the considered statistical units either. In general the adjustments of weights can be effected using calibration methods. The calibration methods consist in adapting the estimated sample distribution of some auxiliary variables to the distribution of the same variables known from outside sources.
In general calibration methods consent to attenuate the distortion due to the phenomenon of totally missing answers, to ensure greater efficiency of sample estimates, to attenuate, whether it is present, the distortion induced by the use of incorrect selection procedures. The data necessary for the adoption of a method of calibration are external to the sample and relative to the distribution of the population for the employed variables. In substance the base weight is multiplied by a multiplicative factor. When the population is divided according to the mode of one or more characters and the joint distribution of these characters is known, that is the number of units of the population belonging to the various dominions defined by the combination of the modes of the various characters is known, we can proceed to an ordinary post-stratification assuming as post-strata the defined dominions.
In substance this is a reweighting based on the known quantity inherent in the population of interest (population weighting adjustments).
Alternatively if the marginal distributions are known, through an iterative process it is possible that the estimates of the marginal frequencies of the modality of the single characters (that is the sum of the weights of the units that represent that modality) be equal to the respective known values. The method introduced by Harora and Brackstone (1977), is known as *raking ratio estimation.*
Furthermore, calibration weighting provides an important class of technique for the efficient combination of data sources. If, from one or more sources, information sufficiently accurate to be used as a benchmark is available on the values of the parameter to be estimated, than sample data can be used to obtain estimates of the quantities of interest, ensuring that the inferential procedure returns results that are close to the reference value.

6. Sensitivity analysis

The application of sensitivity analysis is central for the quality of the simulation especially when models are used for making decision having a large social and economic impact. Methods capable of testing the robustness and relevance of model-based analysis can be considered one of the constituent element of the model building. Sensitivity analysis is a study of how the uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input. It allows to analyse the impact of different factors on estimates. Furthermore, it helps to elucidate the impact of different model structures. Sensitivity analysis can also be used to evaluate which subset of input factors accounts for most of the output variance and in what percentage. Since we are interested to study the output variability, the variance-based methods, also known as *importance measures or sensitivity indices*, can be applied. These methods provide a factor-based decomposition of the output variance to describe output variability. In this way it is possible to identify the most important factor of the model building that would lead to the greatest reduction in the variance.
At the same time the sensitivity analysis can be applied in the various phases of construction of the data-set for micro-simulation. In fact, for the reconstruction of the missing data, sensitivity analysis was performed. The value of the missing datum has been estimated by applying a parametric method based on multiple regression model. Two explanatory variables selected after the analysis of the Pearson Correlation Coefficients and two dummies define the model. This model has been tested considering alternatively a model with all the variables in a linear form and a logarithmic model with two logarithmic variables and two dummies. Statistic tests identify the *best* models and sensitivity analysis methodologies calculate the data quality indicators in the two different approaches adopted to 'model incomplete data' (*single imputation* and *multiple imputation)*. Furthermore hypotheses regarding distortion problems due to non sampling errors are tested and alternative estimation are obtained through single and multiple imputation techniques (cf. *EC-JRC, ISTAT 2003*) . Sensitivity analysis is not only a prerequisite to evaluate uncertainties in the input variables and model parameters, but it helps to establish the better integration process and the definition of the micro-simulation models.

7. Conclusions

The development of economic analysis at micro level (micro-simulation, impact analysis of policies, systemic analysis and indicator construction) can be carried out by joining two or more different sources in order to obtain one single database with more information. The integration process has encounters various problems of different nature and requires preliminary actions such as: the definition of economic objectives, the choice and the analysis of the sources available and necessary for the attainment of the economic goals in question. The attainment of economic objectives is strongly conditioned by the availability and quality of information, as well as by the sample and non-sample nature of the involved data sources. Even the choice of integration methodologies to be used depends on the above mentioned factors. From a methodological point of view the statistical analysis, the data treatment, the application of different integration techniques, the estimate of parameters and the investigation of statistical quality indicators for multi-source databases are complex phases of the integration process. For instance, for the completion of the whole integration process of the Irpeg module, partially analysed in this work, we need also to reconstruct integrated

data-set whit longitudinal information. At a micro level, while for the LE survey a reconstruction of longitudinal information is possible, for the SME survey, since there is not a panel sample, it is not possible. For the latter the reconstruction of longitudinal information relative to some variables is possible by constructing a panel of statistical units with commercial accounts data available for the previous year. Furthermore the couples of records in common between the SME survey and the commercial accounts for the two years can be used to calculate adjustment factors.

Analysis in this paper has tried to describe the main steps and methodological issues of the integration process and to stress its complexity. The process of integration is still under construction which range from census and survey to administrative (including fiscal) data. In this instance, the integration of all available information on enterprises into one multi-sources database realised in the DIECOFIS project gives to the system high potentialities and opportunities in terms of economic analysis. It enables to solve the data requirements of the micro-simulation model for the estimation of the effects of fiscal policies on enterprises performance but also more economic issues can be investigated and analysed in a very detailed perspective through new demand driven methods. The use of a systematised and integrated system makes it possible to create new micro founded indicators that are more appropriate to describe different economic systems and to understand their systemic strength and weakness.

References

Abbate C. (1997), " Completeness of Information and Imputation From Donor With Minimum Mixed Distance ", Quaderni di Ricerca ISTAT, n. 4/1997, pp. 68-102.

Alworth, J. S., Castellucci, L. (1993), Chap. 6, Italy, in Jergenson D. W.- Landon, R. (eds), Tax Reform and the Cost of Capital. An International Comparison, The Brookings Institution, Washington D.C.

Ballin, M., Falorsi, P.D., Falorsi, S., Pallara, S. (2000) Il trattamento delle mancate risposte totali nelle indagini ISTAT sulle Famiglie e sulle Imprese (Analysis of total non-response in ISTAT Surveys of Families and Firms) – ISTAT Methodological Studies.

Bardazzi, R., F., Pazienza, M.G., Parisi, V. (2003), The Effects of the Italian Tax Reform on Corporations: a Microsimulation Approach. Available on the website http://www.istat.it/diecofis.

Bardazzi, R., Gastaldi, F., Pazienza, M.G. (2002), The IRAP module, Deliverable 5.2 of Diecofis Project. University of Florence. Available on the website http://www.istat.it/diecofis.

Bardazzi, R., Gastaldi, F., Pazienza, M.G. (2003), The Social Contribution module, Deliverable 5.1 of Diecofis Project. University of Florence. Available on the website http://www.istat.it/diecofis.

Black, J. (2001) Changes in Sampling Units in Surveys of Businesses. In: 2001 FCSM Research Conference Papers, US Census Bureau.

Bontempi, M.E., Giannini, S., Guerra, M.C., Tiraferri, A. (2001), Incentivi agli investimenti e tassazione del reddito di impresa: una valutazione delle recenti innovazioni normative, mimeo, available on the website http://www.capp.unimo.it.

Bordignon, M., Giannini, S., Panteghini, P. (2001), "Reforming Business Taxation: Lessons from Italy?", in International Tax and Public Finance, vol. 8, n. 2.

Bosi P., Guerra M.C. (2002), I tributi nell'economia italiana, Il Mulino, Bologna.

Brick J.M., Kalton G. (1996), "Handling Missing Data in Survey Research", Statistical Methods in Medical Research, vol. 5, pp. 215-238.

Brackstone, G. (1999), "Managing Data Quality in a Statistical Agency", Survey Methodology, December 1999, vol. 25, pp. 139-149.

Castellucci, L., Coromaldi, M., Parisi, V., Perlini, L., Zoli, M., (2002), Report describing country IT Corporate Tax Model and methodology, Deliverable 6.1 of Diecofis Project, University of Rome Tor Vergata. Available on the website http://www.istat.it/diecofis.

Denk M., Oropallo F. (2002), Overview of the Issues in Longitudinal and Cross-Sectional Multi-Source Databases, www.istat.it/diecofis.

Denk M., Inglese F., Calza M.G. (2003) , Assessment of different approaches for the integration of ample surveys, http://www.istat.it/diecofis/deliverable_list.htm..

Denk, M. (2002) Statistical Data Combination: A Metadata Framework for Record Linkage Procedures. Dissertation Thesis, Dept. of Statistics, University of Vienna.

Denk, M., Froeschl, K.A. (2000) The IDARESA Data Mediation Architecture for Statistical Aggregates. Research in Official Statistics 3 (1), 7–38.

Deville J. C., Särndal C. E. (1992), "Calibration Estimators in Survey Sampling", Journal of the American Statistical Association, vol. 87, pp. 376-382.

EC-JRC and ISTAT (2003), Software analysis, Development of methodologies and of a software for the measurement of statistical quality, and for comparing the robustness of alternative multi-source,integrated database, http://www.istat.it/diecofis/deliverable_list.htm.

D'Orazio, M., Di Zio, M., Scanu, M. (2001) Statistical Matching: a tool for integration data in National Statistical Institutes – paper ISTAT for NTTS 2001 – ETK 2001 – Crete Conference 18-22 June 2001 – Eurostat, JRC(ISIS).

Eurostat (1999) Use of Administrative Sources for Business Statistic Purposes: Handbook on Good Practices – Theme 4 (Industry, Trade and Services), Eurostat Edition.

Eurostat (1999), Model quality report in business statistics, Eurostat Working Group "Assessment of quality in business statistics".
Falorsi, P.D., Falorsi, S. (1995) Un metodo di stima generalizzato per le indagini sulle famiglie e sulle imprese (A Generalized Estimation Method for Surveys of Families and Firms). Quaderni CON PRI, University of Bologna.
FCSM – Federal Committee on Statistical Methodology (1980) Report on Exact and Statistical Matching Techniques, Statistical Policy Working Paper 5, Washington, DC: U.S. Department of Commerce.
Giovannini, E., Sorce, A. (2001) Integration of Statistical (survey) data with registers (administrative) data. Paper contributed to the Meeting on the Management of Statistical Information Technology 2001.
Informer SA (2003a) Code for user interface http://www.istat.it/diecofis/deliverable_list.htm.
Informer.SA,(2003b),Software-User.Manual http://www.istat.it/diecofis/deliverable_list.htm.
Inmon, W. H. – "Data Marts and Data Warehouse: Information Architecture for the Millenium" – Informix Corporation.
ISTAT (2000) I risultati economici delle medio-grandi imprese Anni 1998-99 (Economic Outcomes of Medium-Large Size Enterprises) - Statistiche in breve - July 2000 (LE Survey) - http://www.istat.it/Imprese/Struttura-/index.htm.
Kadane, J.B. (1978) Some Statistical Problems in Merging Data Files. In: 1978 Compendium of Tax Research, US Dept. of the Treasury, 159–171. (Reprinted in Journal of Official Statistics 17 (3), 423–433.
Kalton G., Kasprzik D. (1986), "The treatment of missing survey data", Survey Methodology, vol. 12, n. 1, pp. 1-16.
Kalton G., Kasprzik D. (1982), "Imputing for missing survey responses", Proceedings of the Section on Survey Research Methods, American Statistical Association.
Kamakura W. A., Wedel M. (1997), "Statistical Data Fusion for Cross-Tabulation", Journal of Marketing Research, vol. 34, pp. 485-498.
Kovar, J.G., Whitridge, P.J. (1995) Imputation of Business Survey Data. In: Cox et al. (eds.), Business Survey Methods, New York: J. Wiley.
Lenz, H.-J. (1998) Multi-Data Sources and Data Fusion. In: Proc. New Techniques and Technologies for Statistics (NTTS) 1998, EUROSTAT, 139–146.
Little R. J. A, Rubin D. B. (1987), "Statistical Analysis with Missing Data", Wiley & Sons, New York .
Malvestuto, F.M. (1991) Data Integration in Statistical Databases. In: Michalewicz (ed.), Statistical and Scientific Databases, Chichester: Ellis Horwood, 201–232.
Oropallo F. (2002), Conceptual, logical and physical model of datamarts deliverable 2.1 - http://www.istat.it/diecofis/deliverable_list.htm.
Oropallo F., Skalbania, D. (2003), Concept of IT framework issues and development of software for the creation of a multi-source data base - Analysis of the Software – deliverable 2.2 - http://www.istat.it/diecofis/deliverable_list.htm.
Oropallo F., Caruso E., (2003), The management of DIECOFIS database - deliverable 2.3 - http://www.istat.it/diecofis/deliverable_list.htm.
Paass G. (1986), "Statistical match: Evaluation of existing procedures and improvements by using additional information". G.H.Orcutt and H.Quinke (eds) Microanalytic Simulation Models to Support Social and Financial Policy. Amsterdam: ElsevierScience, 1986, pp. 401-422.
Roberti P., Oropallo F., Inglese F., Lo Cascio L., de Martinis G. (2003) - Towards a Systemic Analysis of Italian Industrial Texture Review "Industria" 4/2002 , Il Mulino – November 2002
Roberti P., Oropallo F. (forthcoming) - Composite Indicators for the Measurement of Economic Performance - Productivity, Competitiveness and the New Information Economy - Business, Systemic and Measurement Issues - NESIS FP5 - ISTAT – Rome - June 26, 2003
Renssen R. H. (1998), "Use of Statistical Matching Techniques in Calibration Estimation", Survey Methodology, vol. 24, n. 2, pp. 171-183
Rodgers, W.L. (1984) , An Evaluation of Statistical Matching. Journal of Business and Economic Statistics 2, 91-102.

Rubin D. B. (1987), Multiple Imputation for Nonresponse in Surveys, Wiley & Sons, New York

Rubin, D.B., Belin, T.R. (1991) Recent Developments in Calibrating Error Rates for Computer Matching. In: Proc. 7th Annual Research Conference, Washington, DC: U.S. Bureau of the Census, 657–668.

Ruggles, N., Ruggles, R. (1974) A Strategy for Merging and Matching Microdata Sets. Annals of Economic and Social Measurement 3 (2), 353–372.

Scheuren, F., Winkler, W.E. (1993) Regression Analysis of Data Files that are Computer Matched. Survey Methodology 19, 39–58.

Scheuren, F., Winkler, W.E. (1997) Regression Analysis of Data Files that are Computer Matched II. Survey Methodology 23, 157–165.

Schafer J.L., Olsen M. .K. (1998), "Multiple imputation for multivariate missing-data problems: a data analyst's perspective", Multivariate Behavioral Research, vol. 33, pp. 545-571.

Singh A. C., Mohl C. A. (1996), "Understanding Calibration Estimators in Survey Sampling", Survey Methodology, vol. 22, n.2, pp. 107-115.

Singh, A.C, Mantel, H.J., Kinack, M.D., Rowe, G. (1993) Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption, Survey Methodology, June 1993- vol 19, No.1 pp. 59-79 - Statistics Canada.

Winkler W. E. (1995), "Matching and Record Linkage", in B. G. Cox et al. (ed.), Business Survey Methods, Wiley & Sons, New York, pp. 920-935 (355-384).

Winkler, W.E. (1993) Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage. In: ASA, Proc. Section on Survey Research Methods, 274–279. Also available as RR93-12, Washington, DC: U.S. Bureau of the Census, Statistical Research Division, http://www.census.gov/srd/www/byyear.html.