SURVEYS ON SMALL BUSINESS AND ACCURACY OF NATIONAL ACCOUNTS IN THE SERVICES SECTOR

Bruno Bracalente, University of Perugia - **Claudio Pascarella**, ISTAT, Rome
Bruno Bracalente, Dipartimento di Scienze Statistiche, Università di Perugia, C.P. 1315 / Succ.1
06100 Perugia (Italy)

KEY WORDS: accuracy, survey, errors

1. INTRODUCTION

In developed statistical systems, information for national accounts mainly comes from enterprise and establishment records through a plurality of surveys on businesses of different size. The accuracy of national accounts aggregates much depends on the quality of data collected by means of said surveys.

It is well known that information abaout sensitive topics such as the value of production and income of small business collected by surveys are typically affected by response and coverage errors (1). In the usual case of list sampling, coverage errors come from the extreme difficulty in maintaining list frames of small businesses because of their peculiar dynamic nature. Response errors mainly depend on the great incidence of under-reporting. When asked about income, finance and other similar topics, small business respondents behave in about the same way that household respondents behave (Cox, 1989; Sudman, 1989). As household income is under-reported, so we must expect of the small business income. Moreover, smaller businesses are not always able to provide the full information requested by the national accounts, so that imputation procedures to remedy lack of information must be extensively employed.

In the market services sector small business usually accounts for the very large majority of all economic aggregates. Actually, the greatest part of employment, value of production, income etc. often belongs to micro business with few employees. Mainly because of this extreme pulverization of the enterprise system, the estimates of economic aggregates of the services sector (at current price), can be seriously affected by errors.

The paper analyzes the main sources of error that affect estimates of the value added in the market services sector in Italy. It briefly describes the procedure used to estimate value added in Italian national accounts and the role sample surveys on small business play on it (par. 2).

The principal shortcomings of this last survey are analyzed (par. 3) along with other potential sources for error commonplace in the evaluation of value added (par. 4). The estimate of the main components of sampling and non-sampling errors is conducted by making reference to the case of business services (par. 5).

Some final remarks on possible alternative techniques in gathering information for national accounts purpose conclude the paper (par. 6).

2. THE ESTIMATE OF VALUE ADDED AND SURVEYS ON SMALL BUSINESS IN ITALY

The estimation procedure

In Italian national accounts, the procedure used to obtain the estimates of value added is based on: i) estimates of value added per employee, in the services sector mainly performed by sample surveys on small business; ii) estimates of employment, used for the expansion of the sample estimates to the target population and obtained by integration of available information drawn from censuses of population and industry and current surveys on households and businesses (2). More specifically, the calculation of the initial estimates of the value added for a branch j is obtained by using the following formula:

$$Y_j = \sum_h V_{jh} L_{jh} + Y_j' \qquad (1)$$

where $h = 1, ..., 4$ refers to the four categories of size of business in terms of employees (1-9, 10-19, 20-49, 50 plus); V_{jh} and L_{jh} respectively represent the value added per employee and employment in category h; Y_j' represents the possible component of the value added independently estimated.

Three different surveys are currently used to assess the value added per employee V_{jh}: a) an annual survey on all the businesses with at least 20 employees; b) an annual sample survey on businesses with between 10 and 19 employees;

c) a pluriennial sample survey on the businesses with between 1 and 9 employees.

Employment evaluation

The following general formula is used to estimate employment according to sector and size of business:

$$L_{jh} = r_{jh}R_{jh} + i_{jh}I_{jh} + n_{jh}N_{jh} + s_{jh}S_{jh} + d_{jh}D_{jh}$$

where r, i, n, s and d are full-time equivalent coefficients for the corresponding categories of "work positions" R (*regular*), I (*irregular*), N (*undeclared employees*), S (*non-resident foreigners*), D (*second job*).

For the years in which the census was carried out, the main components of employment (i.e. R, I and part of D) directly derive from comparison between the two censuses of population and industry.

With regard to the years between censuses such components are estimated by separately up-dating employment figures on the supply side and on the demand side.

The first up-dating is done by means of sample surveys on the labor force. The second through surveys done on businesses.

The role of surveys on small business

Percentage distributions of employment and value added by size of business in the market services sector in Italy (1992) are the following:

Number of emplyees	Employment	Value added
1-9	68.7	57.3
10-19	10.6	11.0
20-49	3.0	3.6
50 plus	17.7	28.1
Total	100.0	100.0

The first category of size of business in formula (1), i.e. between 1 and 9 employees, is definitely the most important. It represents 68.7% of employment and 57.3% of the value added. In the business services sector in particular 80% of employment and 65.3% of the value added belongs to the category of firms of this size (almost 70% of employment actually belongs to micro businesses numbering between 1 and 5 employees).

National accounts estimates greatly depend on the qualitative characteristics of the sample survey done on small business, this being the one that presents the most serious shortcomings. Furthermore, estimates are affected both if these surveys are not carried out and an indirect evaluation of characteristic parameters is performed, and by the methods used to assess and up-date employment figures.

3. MAIN SOURCES OF ERRORS IN THE SURVEYS ON SMALL BUSINESS

The principal sources of errors in the surveys on small business are: a) difficulty in correctly identifying the population of the firms; b) the widespread presence of the phenomenon of under-reporting, this being typical of small business.

Incompleteness of the business register

A business register for statistical purposes is a centralized list of individual economic units used by every statistical survey. It is made up of identification and structural information, such as name and address, industry code, ownership code, number of employees or other size indicator, link to related establishments, etc. Implementation and up-dating of business registers are generally based on multiple sources of information: one or more administrative registers, census and sampling surveys, etc.

For small business, maintaining an up-to-date register is, however, a very complicated matter, mainly because of the high rates of natality and mortality of the component units. As a consequence, in some countries a form of "cutoff register" has been devised, that is a business register in which no units below a specified size are considered.

In the Italian statistical system, the central register of economic units only regards those with at least 10 employees. The register for businesses with between 1 and 9 employees is made up of the list compiled from the decennial census on industry. It is not up-dated during the period between censuses. The register is most incomplete especially with regard to the market services sector where natality and mortality of businesses is elevated and for the years furthest away from the census on industry.

An assessment of the importance of bias due to incompleteness of the register has not yet been made. This would require an ad hoc survey to measure the discrepancy between the value added per employee for the population covered by the register and that not covered by the same. An estimate of the relative incidence of the latter on the total would also be necessary. It should however be assumed that the bias is positive. This is because the

businesses excluded from the survey[150q are of a more recent constitution and of generally inferior dimension and thus of lower profitability.

The validation of this hypothesis and, more generally, an approximate measurement of the coverage error will be possible when data gathered by the last census of industry is made available.

Reporting behaviour

The phenomen of under-reporting represents a typical example of response bias due to the deliberate failure of the respondent to report the correct value. Such a reporting behaviour, typical of the self-employed who mainly run small family businesses (Smith, 1986), is common to surveys on households and on businesses. It is responsible for a part of the hidden economy, so that national accounts must be adjusted upwards by certain criteria.

The assessment of said phenomenon carried out by ISTAT is based on so-called indirect methods. It gets its inspiration from the work done by Franz (1985) on the estimates of the hidden economy in Austria. This criterion is based on the hypothesis that costs undergone by the business are declared as a whole. While the response relative to turnover is considered under-reported when the self employed income per-head results as being inferior to that of the employees.

The procedure adopted for correcting turnover and value added is based on the upwards revision of self-employed income per head, increased to the level of that of employees. According to the above criterion, in the market service sector the incidence of small businesses defined as being under-reporting on the total is of 66.7%. The phenomenon is widespread and so the extent of under-reporting is considerable as will be seen further on.

4. OTHER SOURCES OF ERRORS

Up-dating employment figures

As has already been said, sources of error different to those concerning the survey on small business are reflected on the national accounts estimates. One of these regards the procedure used to update employment figures. An accurate evaluation of the virtues of the criteria used in up-dating can be done when the information gathered from the 1991 census is made available. It will thus be possible to recalculate the figures for employment for that year by direct method and these can then be compared with the figures obtained through the indirect up-dating procedure.

However it is already possible to measure the effects on employment estimates which derive from the delay in making information available to the central register of economic units. The assessment of error can be done by comparing employment figures for a given year, these being obtained from information made available in that year, and employment figures for the same year evaluated by using information subsequently made available.

A slight modification on the distribution of employment according to class of employee resulted from experiments done in the market services sector. In any case, the effects in the business service sector turned out to be of major importance: the new estimates are considerably higher both for the class numbering between 1 and 9 employees and especially for the one with 20-49 employees. As we will see further on this also considerably affects the estimates of value added in this sector.

Indirect parameters evaluation

The indirect assessment of value added per employee for years in which there was no survey on small business represents a specific source of error. For these years the said parameter is evaluated by applying the variation observed for the class with 10-19 employees to the value added per employee of the last year for which a survey was done.

The extent of the error associated with this indirect methodology can be evaluated for the years in which a survey exists. In this case both direct and indirect methodologies can be applied and their results compared. This has been done for the business services sector and the results will be illustrated in the following paragraph.

5. FROM SURVEY TO NATIONAL ACCOUNTS: THE CASE OF BUSINESS SERVICES

An empirical evaluation of the main errors identified in the previous paragraph will now be done. The said evaluation regards small firms (1-9 employees) in the Italian business services sector for the last year for which a survey on businesses of this size was conducted (1988). In order to obtain the most complete evaluation of accuracy of the estimates, the empirical analysis was also extended to the measurement of sampling errors (3).

With regard to the evaluation of sampling errors it must be noted that for the part of value added estimated by sample survey (1) represents a separate ratio estimator where the employment constitutes the auxiliary variate assumed to be known. The evaluation of $V_{jh}L_{jh}$ for $h = 1$ is in fact obtained through stratification according to economic activity and post-stratification according to class of employee (1-5 and 6-9 employees) hence:

$$V_{j1}L_{j1} = \sum_k \frac{\hat{y_k}}{\hat{x_k}} L_k$$

where k indicates the general stratum (k=1, ..., K); $\hat{y_k}$ and $\hat{x_k}$ are the stratum estimates of the total value added and employees respectively; L_k represents stratum employment figures assumed to be known.

As we know the bias of the ratio estimate is negligible if the sample size is sufficiently high. But in stratified sampling the sample size must be high in every stratum. Thus for relatively small samples with a large number of strata, as in the case in question, bias may not be negligible (Cochran, 1977).

Non-sampling errors

With $V_k^{(d)}$, $V_k^{(r)}$ and $V_k^{(c)}$ we will respectively indicate the value added per employee of the stratum k declared by the respondents to the survey, re-evaluated for under-reporting (according to the so-called "Franz method") and corrected so as to take also into account the problems of coverage. Moreover with L_k and $L_k^{(c)}$ we will respectively indicate the provisional estimate of employment figures and those corrected to eliminate errors arising from delay in the up-dating of the central register of firms.

The national accounts estimate of the part of value added concerning small businesses (with 1-9 employees) can be seen as a sum of a "base estimate" and a series of corrections done to eliminate various sources of errors, as in the following formula:

$$Y = \sum_k V_k^{(c)} L_k^{(c)} = Y_0 + E_r + E_c + E_e$$

where:
$Y_0 = \sum_k V_k^{(d)} L_k$ (base estimate);
$E_r = \sum_k (V_k^{(r)} - V_k^{(d)}) L_k$ (response error);
$E_c = \sum_k (V_k^{(c)} - V_k^{(r)}) L_k$ (coverage error);
$E_e = \sum_k (L_k^{(c)} - L_k) V_k^{(c)}$ (expansion error).

If with $V_k^{(u)}$ we indicate the estimate of value added per employee done, in the absence of a survey, through indirect up-dating of the said parameter we can measure the corresponding error with the following formula:
$E_u = \sum_k (V_k^{(u)} - V_k^{(r)}) L_k^{(c)}$ (up-dating error).

As has already been said, due to lack of information it was not possible to evaluate the component E_c in the analysis done in this paper with reference to business services. With regard to the other components the results obtained are given in the following table:

Components of value added	Absolute value (billion lire)	Percentage on total value added
Base estimate (Y_0)	37,127	77.9
Response error (E_r)	8,186	17.2
Expansion error (E_e)	2,358	4.9
Total (Y)	47,671	100.0
Up-dating error (E_u)	5,114	10.7

The base estimate got from the sample survey on small business represents 77.9% of the corresponding national accounts aggregate. The difference between the latter and the base estimate is mainly due to under-reporting: the response error affects 17.2% of the national account estimate. We must also underline the fact that the estimates are highly sensitive to the use of alternative correction techniques of response bias. If the phenomenon of under-reporting is assimilated to item non-response and the corresponding correction is done using a regression imputation technique, response bias has a much greater influence on the result (see Pisani and Viviani, 1993) (4).

An error of lesser importance is attributable to the use of a provisional estimate of employment figures as an expansion factor: the value added turns out to be under-estimated by 2,358 billion lire (4.9% of the total value added).

The error arising from the use of indirect methodology for the up-dating of value added per employee in the years when the survey was not available reveals itself to be of major importance. In fact by indirectly up-dating the results of the survey relative to 1986 the estimate of value added in 1988 is 5,114 billion lire higher than the present one, i.e. an over-evaluation of 10.7%.

Standard error and sampling bias

The estimates of variable error and sampling bias are the following:

	Absolute value (billion lire)	Percentage on total value added
Sampling bias	-43.0	-0.08
Standard error	1,174	2.69

With regard to the case under study, the bias of the ratio estimate is totally negligible: when related to the total value added, this bias does not exceed 1 per thousand. The variable error, measured by the standard error, reveals itself as being quite moderate, at least when compared to the entity of non-sampling errors. Related to the estimate of total value added, the standard error determines a coefficient of variation of 2.69%.

6. CONCLUDING REMARKS

The accuracy of national accounts estimates in the market services sector depends on the characteristics of the sector, which is largely comprised of small business. In the Italian case study which has been analyzed in this paper, the aforementioned characteristics are responsable for: *coverage errors*, due to lack of an up-to-date central register of small economic units in inter-censual years and because of a potential difference between frame population and target population; *response errors*, mainly due to a high incidence of under-reporting, a widespread phenomenon among small businesses. Italian national accounts estimates are also affected by *expansion errors*, because of problems in updating of employment estimates in inter-censual years. Moreover, the procedure of indirect evaluation of parameters such as value added per employee represents a considerably important source of potential error for the years in which surveys were not conducted on small business.

Sampling errors reveal themselves to be of a negligible (bias) or moderate (variable error) entity, at least when compared with non-sampling errors. The accuracy of estimates, therefore, largely depends on the latter, some of which prove difficult to reduce. This in particular is the case of under-reporting, whose correction at the survey stage is quite difficult. Therefore, it must be treated as an ex-post correction using imputation techniques, whose effectiveness is nevertheless dubious.

Coverage error and (to some extent) expansion error may be reduced by extending the central register of economic units to those with less than 10 employees. Actually, the development of business register was one of the principal recommendation of an international commission established by the Italian Government to analyze the Italian statistical system (see Moser et al., 1983).

Both theoretical and practical suggestions for implementing and up-dating a complete business register for statistical purposes come from numerous experiences of other countries, as well as from Italian experiences at a regional (cfr. Martini, 1993) and national scale, the latter carried out by ISTAT itself and aimed at reducing coverage errors in the 1991 census of industry.

As an alternative, at least multiple frame estimation techniques should be adopted in order to reduce coverage error. A frequently adopted solution is to use an area frame to supplement the list frame. However, the well known inefficiency of area sampling can be (partially) reduced only through accurate stratification of area units, that is generally possible only by means of data gathered by decennial census of industry. Area frames are therefore less efficient when used on dynamic and instable populations such as small businesses.

Any way, the empirical analysis developed in this paper seems to demostrate that the most important source of inaccuracy of national accounts is the very high extent of under-reporting. Data on economic results of small business collected by surveys seems to be of the same order of accuracy of data gathered by fiscal authorities.

So a more radical (and less expensive) alternative should also be considered, that is the systematic use of fiscal data, already available in the administrative archives of government, for national accounts purposes. Under-reporting of economic results should still be the central problem, but could be faced by means of the same techniques used for correcting data gathered by surveys. On the other hand, fiscal data are collected annually, so that the "up-dating error" would disappear, while the coverage problem would be reduced to the economic units practicing total tax evasion.

Of course, the fiscal source refers to institutional units, as enterprises, not to establishments, that is single places of goods and services production, so that it cannot provide the full information necessary to the industry disaggregation of national accounts.

Therefore, the use of fiscal data for national accounts purposes cannot supplant surveys on business. Instead, it can positively contribute to shift

the concern of the latter from the estimation of some aggregates to understanding structural characteristics and possibly economic behaviour of businesses.

NOTES

(1) On quality issues in establishment surveys see, among others, Plewes (1988), Tupek and Copeland (1988).
(2) For a description of Italian national accounts methodology, see Istat (1990).
(3) On the accuracy of national accounts see Novak (1975).
(4) This alternative evaluation of under-reporting gets its inspiration from a study carried out by Pissarides and Weber (1985) to estimate Britain's hidden economy by using data drawn from a family expenditure survey.

REFERENCES

Cochran W. (1977), *Sampling Techniques*, Wiley, New York.

Cox B.G. (1988), "Surveying small business about their finance", *ASA proceedings of the Section on Survey Research Methods*, 553-7.

Franz A. (1985) "Estimates of the hidden economy in Austria on the basis of official statistics" *The Review of Income and Wealth*, n. 4, 325-336.

ISTAT (1990), "La nuova contabilità nazionale", *Annali di Statistica*, IX, 9, Roma.

Martini M. (1993), "Integration of Different Administrative Business Registers for Statistical Purposes" (in this volume).

Moser C. et al. (1983), "Aspetti delle statistiche ufficiali italiane. Esame e proposte", *Annali di Statistica*, ISTAT, Roma.

Novak G.J. (1975), "Reliability criteria for national accounts", *The Review of Income and Wealth*, 21, 3, 323-344.

Petrucci A. e M. Pratesi (1993), "Listing frames and maps in area sample survey on establishments and firms" (in this volume).

Plewes T.J. (1989), "Focusing on Quality in Establishment Surveys", *ASA proceedings of the Section on Survey Research Methods*, 71-78.

Pisani S. and A. Viviani (1993), "Note on the estimation of technological coefficients as a methodology to integrate partial responses given by firms" (in this volume).

Pissarides C.A. and G. Weber (1989) "An expenditure - based estimates of Britain's black economy", *Journal of Public Economics*, 39, 17-32.

Smith S. (1986), *Britain's shadow economy*, Oxford University Press, Oxford.

Sudman S. (1988), Discussion of the paper by B.C. Cox (1988).

Tupek A.R. and K.R. Copeland (1988), "Sample Design and Estimation Practices in Federal Establishment Surveys", *ASA proceedings of the Section on Survey Research Methods*, 298-303.

STATISTICAL ASPECTS OF BUSINESS REGISTERS INTEGRATION

Marco Martini
Scuola di Statistica Università di Milano
Via Visconti di Modrone 21-20122 Milano (Italy)

KEY WORDS: business registers, linkage, errors

FOREWORD

A modern statistical system of enterprises is based on a standard set of definitions and classifications and on a business statistical register of enterprises and local units (BSR). That's why some national statistical institutes created the BSR and the European Community Council has adopted a regulation on "Community Coordination in drawning up business registers for statistical purposes" expanded to all enterprises that operate an economic activity and to the local units that depend on them[1]. In Italy the development of the BSR was preceded by some experiments [2].

The present work aims to point out some statistical aspects of business registers integration. After presenting the business administrative registers that are the source for the Italian BSR (part 1), the statistical problems relating to linkage (part 2) and to the treatment of registration (part 3) and attributes (part 4) errors are examined. Some experimental results are then given.

1. THE BUSINESS ADMINISTRATIVE REGISTRES IN ITALY

The principle sources to establish and periodically update a BSR are the **general** administrative registers, to which almost all enterprises and local units **must** report. Business are required to report information **occasionally,** when changes are made, and **periodically** corresponding to the payment of taxes, social contributions, and fees throughout the year.
In Italy there are six general administrative registers to which enterprises must report[3].

[1]Council Regulation (6,10,1993) on "Community Coordination in drawing up business registers for statistical puposes"(Nanopulos P., 1991, La Statistica Europea delle imprese e il 1992, *Impresa & Stato*, 15, September, 26-30).
See also: ISI (1985), *Bulletin of the International Statistical Institute, 45th Session*, Volume LI, Book 2, Amsterdam; INSEE (1988), *Les sourses statistiques sur les enterprises*, Les collections de l'INSEE, SE, September; Bracalente B. (1991), Il sistema dell'informazione economico statistica in Italia, *Impresa & Stato*, 15, September, 34-38; Biffignandi S. (1991), Microaree territoriali e informazioni statistiche, *Impresa & Stato*, 15, September, 32. Council Regulation (10, 6, 1993) on "Community Coordination in drawing up business registers for statistical puposes". See also: Nanopulos P. (1991), La Statistica Europea delle imprese e il 1992, *Impresa & Stato*, 15, September, 26-30.

[2]Integration experiments with statistical objectives on business data were conducted in the following projects: SICIS - Sistema informativo Censimento Industria e Servizi (ISTAT); ASPO - Archivio Statistico Provinciale dell'Occupazione" (Martini M, Aimetti P., 1989, *Un Archivio delle imprese per l'analisi economica. Fonti, metodi, risultati*, Unioncamere, Milan); ATTIS - Atélier du Traitement Intelligent des Informations Statistiques (Criss-Grenoble, Gruppo CLAS-Milan) included in the DOSES research program (Developement of Statisical Expert Systems) of EUROSTAT; Excelsior (CEE, Unioncamere, Ministero del lavoro) (Aimetti P., 1991, Le fonti amministrative per un sistema statistico delle imprese, in: Martini M. ed., 1991, L'informazione economica per le imprese, *Imprese & Stato*, 15, September, 11-89.

[3]General registers in Italy are:
R1: Companies Registry of Chambers of Commerce, to which all non-agricultural enterprises must report;
R2: Record office of INAIL - Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro - which is mandatory for all enterprises, institutions, or free-lance professionals whose work requires insurance against on-the-job injuries;
R3: Record office of INPS - Istituto Nazionale della Previdenza Sociale - to which registration is required for all enterprises, institutions, and independent professionals with employees;
R4: Taxroll of IVA - Finance Ministry - which includes all enterprises, institutions, and independent professionals who have to pay value-added taxes;

Such registries differ among one another in their **observation field**, considered **attributes**, and data **quality**.

The **observation field** (and the number of records) vary according to the type of entity registered and the payment period (table 1).

Tab. 1 Registers Observation Field

registers units	million 1989	enter-prises	istitu-tions	profes-sionals	local units
R1	3,80	Y	N	N	Y
R2	1,80	Y	Y	Y	Y
R3	1,14	Y	Y	Y	N
R4	5,05	Y	Y	Y	N
R5	2,90	Y	Y	Y	Y
R6	2,70	Y	Y	Y	Y

The entities relevant to a BSR, namely non-agricultural enterprises and their local units, are found solely in R1, and somewhat in R2. Only enterprises are found in R3 and R4, and only local units are included in R5 and R6.

The various registers differentiate themselves by the presence or absence of items (table 2), activity classifications, recording methods[4] and completeness of the data[5].

Regarding the **quality** of the data one must remember that the statistical register is interested in the items of entities (enterprises or local units) economically active. That is, entities which engage in actual productive activity and not those that only exist on paper. From this point of view, every business register is characterized by three types of **errors.**

E1: non-registration of active entities due to non-compliance, delays, etc.;

E2: registration of non-active entities, like those with only a legal existence, that have ceased or suspended operation, or have not yet begun operation;

E3: incorrect values of items due to delays or to carelessness in reporting or verification.

Because of the differences, incompleteness, registration and items errors no administrative register can be directly used for statistical purposes.

The statistical problems of BSR building entail the measurement of such errors, the estimation of the effective size of the universe of the entities included in the BSR's field of observation, and of its structure according to the items X1-X6.

Tab 2 Considered attributes

	Enterprises				Local Units			
Attributes	R1	R2	R3	R4	R1	R2	R5	R6
Y1 Reg code	Y	Y	Y	N	Y	Y	Y	Y
Y2 Fiscal code	p	p	p	Y	p	p	N	N
Y3 Name	Y	Y	Y	Y	Y	Y	Y	Y
Y4 Telephone	p	N	N	N	p	N	Y	N
X1 Legal form	Y	Y	Y	Y	Y	Y	N	N
X2 Address	Y	Y	Y	Y	Y	Y	Y	Y
X3 Activity	Y	Y	Y	Y	Y	Y	Y	Y
X4 Employees	p	p	Y	N	p	p	N	N
X5 Self empl.	p	p	N	N	p	p	N	N
X6 Turnover	N	N	N	Y	N	N	N	N

(p: partial data)

Such measurements are made possible by applying the appropriate statistical models to the results of the linkage between entities.

2. STATISTICAL ASPECTS OF THE LINKAGE

Linkage between entities involves the comparison of the N_r records of the file-register r with the N_s records of the file register s (r,s=R1,..., R6). Three types of linkages must be applied.

i) The **code linkage** is performed on the identification items expressed in code that are applied in more than one register (fiscal code in R1, R2, R3, R4 and telephone number in R1, R5). This requires only simple code matching programs but couples only some of the entities due to the absence and errors in each register relating to such codes.

R5: Record office of SEAT - Società Elenchi Ufficiali Abbonati al Telefono- which includes all economic entities that have telephones;

R6: Record office of ENEL - Ente Nazionale per l'Energia Elettrica - which enlists all commercial users of electricity.

[4]The addresses (province, community, street, municipal number), the economic activity and the legal form are written clearly in some registers, partially in others, and in still others they are coded.

[5]The fiscal code is present in 100% of the cases in R4, while missing in R5 and R6; the telephone number has 100% completeness in R5 but only partial in R1; the number of employees is present for 99% of R3, 70-80% for R2, and 40-50% for R1, while missing altogether in the others; the number of independent contractors is present for 40-50% of R1 and R2, and absent from all others; corporate revenues is present only in R4.

ii) The **alphabetic linkage** is performed on items expressed in words, such as the name and address and, eventually, on codes partially incorrect. The process calls for complex programs that compare non standard texts, words, letters, and their combinations. The alphabetic linkage presents further statistical problems concerning the determination of the degree of likeness between different items and the appropriate measure of direct distance $d(u_r,v_s)$ between pairs of entities (u_r belonging to the register r and v_s to the register s) that can compare every entry of each register to all those of the others. Entities are considered linked if their distance falls with in a prescribed threshold S.

iii) The **multilateral linkage** is based on the concept of transitivity. If u_r is linked to v_s because $d(u_r,v_s) \leq S$, and v_s is linked to z_t (entered in the register t) because $d(v_s,z_t) \leq S$, then u_r is also linked with z_t with an indirect distance given by:

$$d'(u_r,z_t) = \max \{d(u_r,v_s),d(v_s,z_t)\}.$$

Table 3 shows the percentages of linkage obtained in different phases, for different pairs of registers through the "Superlinker" procedure[6] applied to an experimental area.

Tab 3. Matching among Business Registers

Linkage type	R1, R2	R1, R3	R1, R5	R2, R3	R2, R5	R3, R5
code :						
-fiscal code	58.9	71,6	-	64.2	-	-
-telephone	-	-	70.4	-	-	-
alphabetical	24.4	11.4	13.1	20.3	68.7	72.0
multilateral	16.7	17.0	16.6	15.5	31.3	28.0
total	100.	100.	100.	100.	100.	100.

Through the appropriate statistical clustering techniques one can construct clusters form entities directly or indirectly linked, and entered in two or more registers. A certain number of entities will remain unlinked.

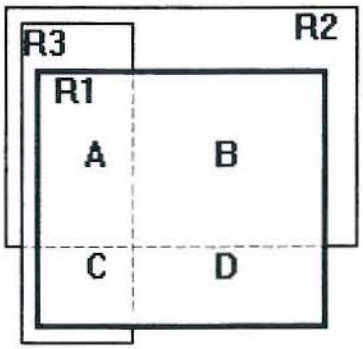
[6]"Superlinker" is a procedure developed by Gruppo CLAS of Milan, and is composed of an optimized series of utilities that perform different types of linkage. In the cited context, the research program DOSES (Development of Statistical Expert Systems) of Eurostat, the possibility is being considered of translating "Superlinker" into an artificial intelligence-based application.

3.STATISTICAL ASPECTS OF THE TREATMENT OF REGISTRATION ERRORS

The most relevant statistical problems concern the treatment of the errors E1, E2, and E3 defined in par. 1, and the estimation of the unknown number of enterprises and local units comprised in the BSR's field of observation.

In order to simplify the discussion, and without loss of generality, only the case involving enterprises included in R1, R2, and R3 will be considered.

The BSR's observation field is subdivided into four subfields A, B, C and D:



The unknown numbres N_A, N_B, N_C, and N_D of enterprises belonging to each subfield, respectively, must be estimated starting from the numbers of enterprises coupled by the linkage:

enterprises registred in:	belonging to the subfields: A	B	C	D
R1, R2, R3:	a_{123}	-	-	-
R1, R2:	a_{12}	b_{12}	-	-
R1, R3:	a_{13}	-	c_{13}	-
R2, R3:	a_{23}	-	-	-
R1:	a_1	b_1	c_1	d_1
R2:	a_2	b_2		-
R3:	a_3	-	c_3	-

given that every register r (r=R1, R2, R3) is characterized by the mutually independent probabilities:

ε_r that an active enterprise eligible for register r is not registered (an E1 type error);

α_r (β_r, γ_r, δ_r) that an enterprise belonging to the subfield A (B, C, D), registered in register r, is not active (an E2 type error).

In particular, R1 is characterized by the probabilities: α_1, β_1, γ_1, and δ_1; R2 by the probabilities: α_2 and β_2; R3 by the probabilities: α_3 and γ_3.

Under these hypotheses, for example, the a_{123} enterprises registered in all three registers and

belonging to the subfield A also include some inactive enterprises and, since the probabilities α_r are mutually independent, there will be: $\alpha_1\alpha_2\alpha_3\ a_{123}$ of them. The $(1-\alpha_1\alpha_2\alpha_3)\ a_{123}$ active enterprises, registered in all three registers, will equal the unknown N_A multiplied by the product of the three independent probabilities $(1-\varepsilon_1)(1-\varepsilon_2)(1-\varepsilon_3)$ that an enterprise is registered in registers R1, R2 and R3 respectively.
Therefore the following equivalences hold for the subfield A:

/A1/ $(1-\alpha_1\alpha_2\alpha_3)\ a_{123}=(1-\varepsilon_1)(1-\varepsilon_2)(1-\varepsilon_3)\ N_A$
/A2/ $(1-\alpha_1\alpha_2)\ a_{12}=(1-\varepsilon_1)(1-\varepsilon_2)\varepsilon_3\ N_A$
/A3/ $(1-\alpha_1\alpha_3)\ a_{13}=(1-\varepsilon_1)\varepsilon_2(1-\varepsilon_3)\ N_A$
/A4/ $(1-\alpha_2\alpha_3)\ a_{23}=\varepsilon_1(1-\varepsilon_2)(1-\varepsilon_3)\ N_A$
/A5/ $(1-\alpha_1)\ a_1=(1-\varepsilon_1)\varepsilon_2\varepsilon_3\ N_A$
/A6/ $(1-\alpha_2)\ a_2=\varepsilon_1(1-\varepsilon_2)\varepsilon_3\ N_A$
/A7/ $(1-\alpha_3)\ a_3=\varepsilon_1\varepsilon_2(1-\varepsilon_3)\ N_A$,

that can be considered as seven equations in the unknowns: α_1, α_2, α_3, ε_1, ε_2, ε_3, and N_A.
From the system /A1/-/A7/, if:

$$M^2=(a_1a_2/a_{12})^2+(a_1a_3/a_{13})^2+(a_2a_3/a_{23})^2+$$
$$+4(a_1a_2a_3/a_{123})-2[(a_1a_2/a_{12})(a_1a_3/a_{13})+$$
$$+(a_1a_2/a_{12})(a_1a_3/a_{13})+(a_1a_3/a_{13})(a_2a_3/a_{23})]$$

and:

$$k_r=1/2\ [a_s/a_{rs}+a_t/a_{rt}-(a_ta_s/a_r)/a_{ts}+M\ 1/a_r]$$
$$(r,s,t=R1,R2,R3;\ r\neq s\neq t)$$

are positive, the following positive solutions are obtained:

$$\varepsilon_r=k_r/(1+k9_r)\ \ (r=R1,R2,R3),$$
$$N_A=M/(\varepsilon_1\varepsilon_2\varepsilon_3),$$
$$\alpha_r=1-\varepsilon_s\varepsilon_t(1-\varepsilon_r)\ N_A/a_r$$
$$(r,s,t=R1,R2,R3;\ r\neq s\neq t).$$

From the following equivalences that hold for the subfield B:

/B1/ $(1-\beta_1\beta_2)b_{12}=(1-\varepsilon_1)(1-\varepsilon_2)N_B$
/B2/ $(1-\beta_1)\ b_1=(1-\varepsilon_1)\varepsilon_2\ N_B$
/B3/ $(1-\beta_2)\ b_2=\varepsilon_1(1-\varepsilon_2)N_B$,

given ε_1 and ε_2, positive solutions for β_1, β_2, and N_B are obtained.
From the following equivalences that holds for the subfield C:

/C1/ $(1-\gamma_1\gamma_3)c_{13}=(1-\varepsilon_1)(1-\varepsilon_3)N_C$
/C2/ $(1-\gamma_1)\ c_1=(1-\varepsilon_1)\varepsilon_3\ N_C$
/C3/ $(1-\gamma_3)\ c_3=\varepsilon_1(1-\varepsilon_3)\ N_C$,

given ε_1 and ε_3, positive solutions for γ_1, γ_3, and N_C are obtained.
And finally for the subfield D, the following applies:

/D1/ $(1-\delta_1)\ d_1=(1-\varepsilon_1)\ N_D$,

from which, given ε_1 and setting $\delta_1=\beta_1\gamma_1/\alpha_1$, yields:

$$N_D=[1/(1-\varepsilon_1)][(1-\delta_1)\ d_1].$$

Analogous procedures can be applied to the analysis of the registration errors for local units.

4. STATISTICAL ASPECTS OF THE TREATMENT OF THE ATTRIBUTE ERRORS

Different values of the item Xh (h=1,...,6), may be assigned to an unit belonging to the BSR and icluded in different business administrative registers in which the Xh is considered: in this case one must choose which value to assign to the unit u in the statistical register.
For example, consider the case of the values x_1, x_2, and x_3 of the item "employees number" present in the three registers R1, R2, and R3 respectively.
Having chosen as a consitency criterion among x_r and x_s, registered in registers r and s, for example, that $|\ln(x_r)-\ln(x_s)|$ is less than some threshold figure, there are:
N the number of units to which the item "number of employees" applies in all three registers;
N_{123} the number of units characterized by consistent values in all three registers;
N_{st}, (s,t=R1,R2,R3; $s\neq t$) the number of units with item values in three registers, but that are consitent in only two, s and t;
λ_r, (r=R1,R2,R3) the probability that the value x_r, registered in r, is incorrect.
Given that the probabilities λ_1, λ_2, and λ_3 are mutually independent, one may write:

$$N_{123}=(1-\lambda_1)(1-\lambda_2)(1-\lambda_3)N,$$
$$N_{st}=\lambda_r(1-\lambda_s)(1-\lambda_t)N,$$
$$(r,s,t=R1,R2,R3;\ r\neq s\neq t),$$

from which comes:

$$\lambda_r=N_{st}/(N_{123}+N_{st})\ (r,s,t=R1,R2,R3;\ r\neq s\neq t)$$

which represents the considered attribute error measurement in the register r.

It is reasonable to assign the unit u in the value which yield the minimum probability of error, denoted λ_u.
Lastly remains the imputation of items which are missing in all registers. This requires the use of different procedures, such as "mean", "hot deck", "cold deck", "regression", "stochastic", "composite" imputation methods[7].
All those methods can be applicated taking into consideration, for every unit u, even the probability π_u that the unit is active, and of the (minimum) probability of error λ_{hu} associated to the item Xh assigned to the same unit u in the BRS.
The "cold deck" method appeals to the use of other secondary sources such as partial or non-mandatory registers, current statistical inquiries on enterprises or local units, or else integrative ad hoc investigation. Even from such sources, thanks to integration, it is possible to estimate the probabilities of error associated with a single attribute with the same methodology outlined above.

FINAL CONSIDERATIONS

From this discussion, it appears clear how the integration of registers by the appropriate statistical methods permits:

i) The estimation of the unknown size and structure of the universe of enterprises and local units belonging to the BSR's observation field (an example of such an estimation, with reference to the structure of the NACE rev.1 sections of economic activity, in a experimental area is reported in table 4);

ii) for every business register or secondary source r, the assignment of:
- the probability of error ε_r that an active entity belonging to the BSR's observation field, is not registered in r,
- the probabilities $\alpha_r, \beta_r, \gamma_r, \ldots$ that a unit registered in r and belonging to the observation subfield A, B, C, ... (respectively) is not active,
- the probability λ_{hr} that the value assigned to the item Xh (h=1,...,6) in r is incorrect;

iii) for every unit u assigned to the BSR and registered in at least one business administrative register, the assignment of:
- the probability π_u that it is active,
- a minimum probability λ_{hu} of committing an error in imputing the value of the item Xh in the statistical register.

Last but not least, the integration of registers makes it possible to perform the above-mentioned estimations annually. This progressively improves the data quality and reduces the cost of the integration or of direct investigation.

[7]See: Chapman D. W. (1976), A survey of non reponse Imputation procedures, *American Statistical Association, Proceedings of the Survey Research Method Section*, 245-251; Ernst L. R. (1980), Variance of the estimated means for several imputation procedures, *American Statistical Association, Proceedings of the Survey Research Methods Section*, 716-720; Kalton G., Kish L. (1981), Two efficient random imputation procedures, *American Statistical Association, Proceedings of the Survey Research Methods Section*, 146- 151; David M. H., Little R. J. A., Samuhel M. E:, Triest R. H. (1986), Altrnative methods for CPS income imputation, *Journal of American Statistical Association*, 81, 29-41; Little R. J. A., Rubin D. B. (1987), *Statistical Analysis with missing data*, J. Wiley.

Tab.4. Enterprises and Local Units, for the NACE rev.1 Sections of Activity[8].

NACE Rev.1 sections	Enterprises					Local Units			
	Statist. Register	Administrative Registers				Statist. Register	Administrative Registers		
		R1	R2	R3	R4		R1	R2	R5
A	4	19	7	-	406	4	19	11	21
B	-	-	-	-	-	-	-	-	-
C	-	-	-	-	2	-	-	-	-
D	195	191	203	86	201	203	198	235	181
E	-	1	-	-	-	-	1	-	2
F	164	172	131	68	144	170	178	148	35
G	464	475	328	148	556	482	519	356	377
H	98	62	93	41	126	102	67	100	101
I	73	71	69	21	79	76	92	74	52
J	38	45	16	15	23	39	45	16	29
K	67	73	37	32	238	70	78	44	183
L	-	-	-	-	2	-	-	-	25
M	10	6	14	5	6	10	6	19	45
N	7	2	8	14	67	7	2	10	46
O	67	109	69	47	101	70	111	70	60
P	1	1	-	-	-	1	1	-	-
missing	-	32	2	-	3	-	32	2	22
	1188	1259	977	477	1954	1234	1349	1085	1179

[8]NACE rev. 1 sections are:
A: Agricolture, hunting and forestry;
B: Fishing;
C: Mining and quarryng;
D: Manufacturing;
E: Electricity, gas and water supply;
F: Construction
G: Wholesale and retail trade and reparing;
H: Hotels and restaurants;
I: Transport, storage and communication;
J: Financial intermediation;
K: Real estate, renting and business activities,
L: Public administration and defence, compulsory social security;
M: Education;
N: Health and social work;
O: Other community and personal service activities;
P: Household services.

ADMINISTRATIVE REGISTERS AND NATIONAL SURVEYS

Silvia Biffignandi, Christine Butti
Dipartimento di Matematica, Statistica, Informatica ed Applicazioni
Università di Bergamo -p.zza Rosate 2-24100 Bergamo - (Italy)

KEY WORDS: small businesses, frame, sample.

1. FOREWORD

This work is included in the range of problems of having at disposal more and more updated information about businesses. Up to now, in fact, information about business entities in Italy has been obtained every ten years through the Census of Industry and Commerce and, periodically, through some specific surveys carried out by interviewing business entities. These surveys have been influenced, among other problems, by the fact that in the intercensus period there is no updated reference to the dimension (number and employees) and to the address of these business entities. Note that the term "business entities" in Italy covers all sectors of economy including areas such as manufacturing, transportation, public administration, services. Strategies in order to improve and increase (both in terms of periodicity and of territorial detail) the quantity of statistical information about business entities are, therefore, essential; this activity involves several research problems. The approach which seems more convenient is the utilization of administrative data for statistical purposes together with the use of administrative records, in the phase of survey sampling as well as in the inference one.

In this work, there are some preliminary remarks on the problem of supplying, creating and updating a frame to be used in the surveys about small businesses (1-9 and 10-19 employees). Our special attention to this range of dimensions is essentially due to a couple of considerations:

a) from the point of view of economic analysis, small business entities are particularly interesting because:

1) they make up, both for number and for employment, more than half of the Italian industrial system;

2) they present differentiated territorial and sectorial dynamics and they characterize the local industrial background quite well;

3) they are extremely dynamic, meaning with this that they can rapidly be made new, they can adapt themselves to new productions, stop or start activities;

b) from the point of view of statistical analysis, of gathering data in particular, they present more consistent problems if compared with larger entities, because: 1) they have high birth and death rates and, therefore, it is much more difficult to have an updated list at disposal; 2) they are internally less organized to supply data, they are tendentially more prejudiced while giving information and afraid that this information is not treated as confidential; for this reason, it is getting more urgent either to find updated survey information, which is to be treated by refined inferential procedures, or to define a project of mutual use of administrative records and surveys, both for the building and updating of the frame and for the gathering of economic data.

With reference to the two surveys carried out by ISTAT (that is, the Italian Central Statistical Office) for the calculation of the value added, one about business entities with 1-9 employees and the other about business entities with 10-19 employees, we want to point out some topics which should be investigated thoroughly. They essentially keep to the frame, to the sampling design and to the inference intended as relationship between sample and target frame in general. The main topics are:

a) the control of the survey coverage, as regards the target frame, with special attention to the possibility to extend the estimates at a more disaggregated territorial level than the one considered at present;

b) the setting up of an updated and adequate frame (obtained with special integrating procedures among databases, conveniently corrected in order to answer statistical objectives rather than administrative objectives only) on which it will be possible to redesign surveys in the future;

c) identification - on the basis of what has come out from point a) - of further methodological remarks about procedures of evaluation, control, adaptation of the data obtained from the survey. This aim could be accomplished through the critical comparison among the frames described from different administrative databases. As a matter of fact, since problem b) needs a long work, recent surveys as well as surveys in progress are not yet based on a frame built on purpose and then updated.

2. SHORT DESCRIPTION OF THE SAMPLING PROCEDURES FOLLOWED BY ISTAT IN THE SAMPLE SURVEYS ABOUT SMALL BUSINESSES.

ISTAT sampling procedures are different in methodology according to the dimensions of the sampled entities. A description of these procedures can

be found in Istat (1991a and 1991b); we will mention here only some basic ideas which can be useful in order to point out the problems which are to be faced; note that in this work we limit our examination to the manufacturing industry of branches 3 (mechanical engineering) and 4 (food, textile industries etc.).
As regards the survey about business entities with less than 10 employees, ISTAT, for the manufacturing sector, takes into consideration only the businesses with 2-9 employees; it is believed, in fact, that very small entities do not come up to the aims of the survey.
The stratified sampling for classes of economic activity ATECO (with two-digit number) and for dimensional classes (2-5, 6-9 employees) is carried out on 37 provinces which are considered significant; the inference is carried out through four territorial divisions (north-eastern, central, southern and insular Italy). Since in the follow-up, the survey coverage will be carried out just on Lombardy, we want to make clear that, in the division concerning north-eastern Italy, the provinces which are included in the sample are Milan, Brescia, Bergamo and Como.
The frame on which the sampling is set out is the Census of Industry and Commerce of 1981; before proceeding to the interviews, a control about the existence of the names is made at the Business Register of the Chambers of Commerce, that is on a database which is run with administrative purposes. Note that the term Business Registers, in Italy, refers to the administrative registers held by the Chamber of Commerce of each province.

As regards the survey on business entities with 10-19 employees, the sample is stratified for territorial division (representative provinces are not selected), for class of economic activity and for class of employees. It is taken out from SIRIO database, that is from a database built by ISTAT on the basis of 1981 Census data and updated through an appropriate survey and analysis of the Business Register as regards businesses which stopped their activity. Due to the fact that the births and deaths of businesses with 10-19 employees are not adequately represented in SIRIO database, ISTAT did not consider to proceed to inferences in the 1988 survey.

3. COMPARISONS BETWEEN DIFFERENT FRAMES AND STRUCTURE OF THE EFFECTIVE SAMPLES OF THE SURVEYS ABOUT THE VALUE ADDED: THE CASE OF A REGION.

3.1 The databases taken into consideration.

As already quoted, the present sampling procedure of these surveys presents two kinds of problems: the first one is concerned with the fact that the reference frame is scarcely updated, the other one - due both to ISTAT choices and to limitations coming from the previous situation - with the estimates which are carried out at the level of group of regions. It is not possible, therefore, to have estimates at a regional or subregional level.
In order to have at disposal indications about the possible procedures, so as to get prospective modifications and integrations of the sampling procedure for regional estimates from the present surveys, we will now examine some data concerning both the sample structure and the universe one, as it appears in Lombardy.
The choice of this region is due to the fact that, for this region, we can dispose, besides SIRIO database (the actual sampling frame) and other administrative databases (for instance the Business Register; INPS, i. e. social security database) of a statistical database based on administrative data, ASPO, the Provincial Statistical Employment Database. ASPO is updated every year. Some comparative and critical remarks on the structure of the industrial system coming out from different databases can, therefore, give the possibility of starting out the research of a methodology able to modify the sampling of this survey, so enabling both the definition of a sample much more corresponding to the characteristics of the frame and the achievement of regional estimates of the economic aggregations.
Before analysing the data, we summarize the characteristics of the databases which we are going to examine: SIRIO, ASPO and INPS.
SIRIO, on whose building and updating characteristics we have already talked about in the previous paragraph, refers only to business entities with more than 10 employees. Its limits from the point of view of a complete and updated frame (we refer to the database concerning the years from 1984 to 1988, since new and more complete updating procedures are being realized) depend on the fact that the data on new-born firms and on businesses which are changing their legal or/and administrative form, are lacking.
ASPO contains every dimensional class: more exactly, it contains "the economically active local units of non-agricultural private business entities". In practice, from the point of view of economic activities, the frame of ASPO corresponds mostly to the Census one, and to the Business Register one. ASPO, however, differs from the Business Register because it only takes into consideration active units, that is units which are economically working. A business entity, in fact, even if it is enrolled in the Business Register and it is

juridically formed, can be not-working economically because it has postponed or suspended its activity; or because, after stopping its activity, it waits some time before dissolving formally and cancelling from the register (For details about the characteristics of ASPO see Martini M., Aimetti P., 1989)

INPS does not refer to local units, like ASPO, but to business entities, as it happens in SIRIO. For the nature itself of the database, only business entities with employees are taken into consideration; according to recurring estimates in terms of productive units, these represent the 35% of the positions reckoned in ASPO database.

3.2 Comparative analysis of frames.

As regards businesses with 10-19 employees, the percentage of units of the effective sample, if compared with the database from which the sample is taken, ranges from 31% to 39% in 7 out of 9 provinces (provincial average on the ATECO 2-digit classification); provinces with a fairly low number of businesses: Sondrio (5%) and Cremona (23%) are an exception.

If we calculate the percentage of sampled businesses according to ASPO frame (to this aim, the number of business entities has been estimated on ASPO database), we get lower percentages, since the ASPO frame gives a higher number of businesses.

Comparing the structure of the frame according to SIRIO, according to estimates for ASPO and according to INPS, we can remark that in all the provinces and in all the classifications of activity (two-figure ATECO), the number of businesses in the strata is sistematically lower in SIRIO (see Table 1). This is particularly strange for INPS database which, as it also considers the business entities with employees, should make up a subgroup of the businesses target frame.

As regards 1988, for instance, the values in SIRIO are almost always lower, for branch 3 and branch 4. This difference is particularly evident for the largest province, Milan, where it can be found + 456 (branch 3) and + 379 (branch 4) in favour of INPS.

As fas as business entities with 2-9 employees are concerned, taken ASPO 1989 as a hypothetical reference frame to which the sampled businesses drawn in 1988 belonged (and "corrected" in advance, in order to refer to business entities rather than to local units), it can be noticed how the sample coverage is variable if compared with two-figure ATECO classification (once the province is fixed): the standard deviation goes from 1.88% (Brescia) to 3.09% (Como). The mean percentage of sampled business entities as to the frame ASPO (a mean calculated first by sectorial classification and next by province) is equal to 3% for branch 3, and to 3.7% for branch 4. As regards Bergamo, the mean of the ratio sample/frame is 1.25% (branch 3) and 3.58% (branch 4) with a maximum of 9.7% in the case of class 42; the variability (std. dev.) is respectively 1.1% and 2.8%. A similar situation is found in the other three provinces which are part of the sample (see Table 2): in general, the percentage of sampled businesses as to the sampling frame is lower in branch 3 than in branch 4, except for Como province where the average of branch 3 is 4.18% and the average of branch 4 is 3.3%, perhaps distorted by a peak of 11.27 % for class 33.

4. FINAL CONSIDERATIONS

The coverage of the effective sample compared with the theoretical one, looks not very high, on the basis of the above-mentioned analysis. Furthermore, let's take into consideration some data concerning the deaths and births of business entities. It is evident that the effective sample represents, instead of a sampling operation, an examination of a portion (or even almost all) of the stable part of the frame. Since this sampling frame is based on a list of business entities drawn eight years before, it contains only the stable portion of the target frame of 1988. In fact, the number of business entities and existing local units does not remain unchanged in the course of time, as it is pointed out in a comparison between ASPO data of 1981 and of 1989 (see table 3).

Besides, if we take into consideration the high number of births and deaths of the business entities with 2-9 employees, it is evident that over a plurennial period, the stock of business entities contains a quite reduced portion of stable businesses.

Let's consider, on this matter, some data concerning Bergamo province: between 1987 and 1989. The 1911 local units which were born in 1987 and survived at least until 1989 , were about 30% of the total amount of 1989 (2-9 employees, branches 3 and 4).

If we assume that even for business entities there are similar birth and death rates, this would imply that the business entities sampled by ISTAT in 1988, among those already existing in 1981, were not representative of the target frame, as long as we do not exclude new businesses from the target population.

From the above considerations it comes out that it is important to make further analyses about the coverage of 1988 sample for the business entities with 2-9 employees and to try and individuate more refined inferential procedures in order to correct this lack.

The analyses made point out that in Italy, with

Table 1
COMPARISON BETWEEN SAMPLE DATA AND REGISTERS DATA

2-9 EMPLOYEES

PROV.	BRANCH	SAMPLE	ASPO	ASPO*	ASPO **	INPS	INPS
			1-9	2-9	bus. 2-9	1-9	2-9
BG	3	56	5461	2990	2891	2757	2003
	4	110	6773	3891	3724	3401	2441
	TOT	166	12234	6881	6615	6158	4444
BS	3	111	7817	4892	4730	3970	2955
	4	146	7838	4301	4199	3604	2578
	TOT	257	15655	9193	8929	7574	5533
CO	3	63	4783	2932	2782	2169	1556
	4	109	8753	4798	4594	3252	2321
	TOT	172	13536	7730	7376	5421	3877
MI	3	314	21377	14603	13299	10303	7664
	4	427	26437	15516	14562	11546	8215
	TOT	741	47814	30119	27861	21849	15879

10-19 EMPLOYEES

PROV.	BRANCH	SAMPLE	SIRIO	INPS	ASPO	ASPO** bus.
BG	3	170	590	656	910	1898
	4	211	930	946	1189	1165
	TOT	381	1520	1602	2099	2063
BS	3	237	821	824	1208	1207
	4	213	1011	1055	1297	1269
	TOT	450	1832	1879	2505	2476
CO	3	161	488	459	717	680
	4	176	630	655	913	884
	TOT	337	1118	1114	1630	1564
CR	3	56	164	193	260	255
	4	70	272	283	302	300
	TOT	126	436	476	562	555
MI	3	700	1855	2311	3976	3725
	4	614	1882	2261	3422	3288
	TOT	1314	3737	4572	7398	7013
MN	3	28	110	123	145	136
	4	83	310	341	421	421
	TOT	111	420	464	566	557
PV	3	77	201	214	282	271
	4	83	252	259	318	306
	TOT	160	453	473	600	577
SO	3	0	18	29	37	33
	4	12	57	59	75	75
	TOT	12	75	88	112	108
VA	3	169	571	522	763	741
	4	232	810	889	1144	1126
	TOT	401	1381	1411	1907	1867

Table 2
COMPARISON BETWEEN SAMPLE DATA AND REGISTERS DATA

PROVINCE: BG

2-9 EMPLOYEES

BRANCH	SAMPLE	ASPO	ASPO*	ASPO **	INPS	INPS
		1-9	2-9	bus. 2-9	1-9	2-9
31	27	2935	1607	1554	1802	1372
32	14	1187	650	628	297	198
33	0	59	32	31	22	17
34	9	757	415	401	438	299
35	0	20	11	11	6	3
36	0	53	29	28	16	13
37	6	450	246	238	176	101
3	56	5461	2990	2891	2757	2003
41	11	630	362	346	446	252
42	6	112	64	61	53	29
43	22	685	394	377	378	304
44	0	121	69	66	72	63
45	20	1585	910	871	787	609
46	23	2029	1166	1116	824	560
47	7	465	267	256	264	202
48	15	748	430	412	458	343
49	6	398	229	219	119	79
4	110	6773	3891	3724	3401	2441
TOT	166	12234	6881	6615	6158	4444

10-19 EMPLOYEES

BRANCH	SAMPLE	SIRIO	INPS	ASPO	ASPO** bus.
31	82	315	448	477	471
32	42	149	79	229	226
33	0	0	2	6	6
34	30	91	102	166	164
35	0	6	2	4	4
36	4	7	10	6	6
37	12	22	13	22	21
3	170	590	656	910	898
41	15	33	29	32	31
42	7	14	14	24	24
43	40	156	148	194	190
44	5	21	19	24	24
45	50	411	437	466	457
46	35	113	112	155	152
47	21	63	60	89	87
48	29	101	116	172	168
49	9	18	11	33	32
4	211	930	946	1189	1165
TOT	381	1520	1602	2099	2063

reference to the surveys about small business entities (in particular, the ones about value added are examined here) in order to improve the reliability of the estimates, a lot of problems must be faced and solved. They concern, among the others, the frame which allows access to the elements of the target population, that is the frame which contains the units to which the probability sampling scheme is applied.

As regards the sampling frame used at present, in fact, it is certainly out-of-date and inadequate. It is, therefore, essential to define a system of frame construction and updating. This result may be achieved through a system which links administrative sources in a different structured way. This requires a consistent amount of work.

Since an adequate frame has not been used in recent surveys, this paper underlines that the results of the already effected surveys should be examined with reference to the realized coverage rate. In particular, we can make out two aspects:

a) analysis of the dimension of the sample at regional and/or subregional level;

b) analysis of the lacks of the sample, from the point of view of the percentage of sampled units compared to the target population.

With reference to the dimension of the sample with reference to regions, the few analyzed data seem to point out that:

- as regards the survey about 1-9 employees, in order to be able to set out estimates more detailed (that is regional estimates) than ISTAT ones (which are, as already mentioned, for groups of regions) the sampling is widely insufficient (we refer to the results given with reference to Lombardy);
- as regards the survey about 10-19 employees, for which ISTAT does not make inferences for the lacks of the sampling frame, the present percentage of sampled units as to the frame is averagely around 21%.

So it seems quite possible to study such forms of integration and correction as to enable a territorial estimate at a more detailed level (regional or provincial), once the problem of the definition af an adequate frame has been solved.

The request of disaggregated territorial information seems to suggest the opportunity of setting out, starting from now, an investigation about the design to follow in order to solve this problem.

As regards the aspect stated in point b) (that is the analysis of the sample from the point of view of the coverage compared to the target population), it should be investigated with reference to the recent surveys: 1) to individuate prospective corrections of estimates; 2) to supply more correct keys to the reading of the results given; 3) to have at disposal remarks which are to be taken into consideration in the future surveys and in the definition of a more adequate sampling frame.

We want to report some considerations:

- as it is well-known, errors of non-observation result from failure to obtain data from parts of the survey population; we can distinguish two different kinds of non-observation errors: nonresponse and non-coverage. Non-coverage refers to failure to obtain observations on some elements selected and designed for the sample; the error, then, can be measured on sampling data.

Unlike non-response, non-coverage denotes failure to include some units, or entire sections, of the defined survey population in the actual operating sampling frame. Because of the actual zero probability of selection for these units, they are in effect excluded from the survey results. This happens without a deliberate exclusion concerning the objectives of the survey, t.i. without an exclusion of the target population.

Whereas non-response can be measured from the sample results, the extent of non-coverage can be estimated only against some checks obtained outside the survey procedure itself. An alternative to obtain information about non-coverage consists of attaching a linking procedure to the sample or to a subsample. This alternative could be evaluated in the case of the surveys examined, with reference to a region at least. As a matter of fact, by attaching sampled data to INPS data and to ASPO data, one could set up information about non-sampling errors.

With reference to the surveys examined, the most important problem concerning the frame imperfection is essentially linked to the undercoverage, as the updating with the new business entities is lacking. This means that the frame gives access only to parts of the target population; the target population is quite different from the frame, then.

This implies problems in obtaining valid inferences and results, above all if we are not in a position to hypothesize that the part of target population which is excluded has characteristics similar to those of the part which is included.

ISTAT, in the survey about business entities with 1-9 employees, takes the following inference procedure.

The businesses sampled can be classified as follows:

a) businesses which belong to the ik stratum (i=class of employees;k=two-digit ATECO sectoral classification), the same stratum in which they have been sampled;

$n_{ik}^{(ik)}$ is their number in each stratum;

b) businesses which belong to the jh stratum (j=class of employees;h=two-digit ATECO sectoral classification),

while they have been sampled in the ik stratum; $n_{ik}^{(jh)}$ is their number in each stratum;

Note that: $\sum_{ij} n_{ik}^{(ih)} = n_{ik}^{(..)}$ i. e., number of businesses which changed stratum.

c) businesses which closed their activity before the survey; $n_{ik}^{(1)}$ represents their number in each stratum ;

d) businesses, which have less than two employees or more than 10; $n_{ik}^{(2)}$ represents their number in each stratum.

Inference is based on the following number of the businesses sampled in the ik stratum:

$$n_{ik} = n_{ik}^{(ik)} + n_{ik}^{(..)} + n_{ik}^{(1)} + n_{ik}^{(2)}$$

Thus, the estimated number of businesses in the ik stratum is:

$$N_{ik}/n_{ik}).\ n_{ik}^{(ik)} = \hat{N}_{ik}^{(ik)};$$

whereas the estimated number of businesses which closed their activity or went out of the ik stratum is:

$$\hat{N}_{ik}^{(1)} + \hat{N}_{ik}^{(2)} = (N_{ik}/n_{ik}).\ (n_{ik}^{(1)} + n_{ik}^{(2)}).$$

In order to take new born businesses into account, their number is estimated as follows:

$$(\hat{N}_{ik}^{(1)} + \hat{N}_{ik}^{(2)})/2$$

and the mean values of the aggregates are hypothesized equal to those actually gathered in the sampled businesses.

We will not discuss here further details of the inference procedure. The information above-mentioned shows that the actual inferential procedure does not admit different economic behaviours between new businesses and business entities already existing in 1981.

In our opinion, it should be deeply investigated how the target population is distributed within the three parts in which it can be subdivided, t.i. :

1) the stable part of population, that is the business entities which keep the dimensional and territorial characteristics for a long period of time (in this particular case from 1981 to 1988),

2) the part of business entities which were more recently born,

3) the part of business entities which have gone out of the target population because they have stopped their activity or have passed to another class of dimension. Since the data are gathered only on the permanent part of businesses, it seems a serious economic hypothesis to apply the data obtained from that part, to the target frame.

Economic analysis is paying much attention to the dynamic and innovative particular behaviours of small business entities and of new business entities above all. Therefore, it may be useful to evaluate the frame imperfections in a more exhaustive way.

Special attention should be devoted to the coverage, (the undercoverage, in particular). The information coming from the economic behaviours specific of new small business entities should be taken into account as much as possible, in order to examine if and how to take the new business entities into account in the estimate phase.

REFERENCES

Istat, Conti economici delle imprese con addetti da 10 a 19, Year 1988, Collana d'Informazione, n. 17, Rome, 1991.

Istat, Conti economici delle imprese con meno di 10 addetti, Year 1986 and 1988 , Collana d' Informazione, n. 45, Rome, 1991.

Kish L., Survey Sampling, USA,1965.

Martini M., Aimetti P., Un archivio delle imprese per l'analisi economica, Unioncamere, 1989.

Saerndal C.E., Swensson B., Wretman J., Model Assisted Survey Sampling, Springer Verlag, 1992.

Paragraphs 1 and 4 are due to Silvia Biffignandi, paragraphs 2 and 3 to Christine Butti.

This work has been developed within a CNR financial support

Table 3:
LOCAL UNITS BORN IN BG PROVINCE IN 1981-89 AND 1987-89

PERIOD	BRANCH	2-9 EMPL.	% ON 1989 TOTAL	10-19 EMPL.	% ON 1989 TOTAL
1981-89	3	1972	66%	433	48%
	4	2334	60%	645	54%
	TOT	4306	63%	1078	51%
1987-89	3	732	24%	109	12%
	4	1179	30%	171	14%
	TOT	1911	28%	280	13%

NOTES TO THE TABLES:

- Aspo data refer to 1989.

SYMBOLS USED IN THE TABLES:

- * = Estimate of the number of local units with 2-9 employees. It is obtained by subtracting the percentage of local units with only one employee (calculated from census data) from the number of local units with 1-9 employees.
- ** = Estimate of the number of businesses (as opposed to local units) obtained by applying the percentage of businesses over local units, at the level of province, one-digit branch and class of employees classifications.
- *** = Estimate of the number of businesses with 2-9 employees. It is obtained by subtracting the percentage of businesses with only one employee (calculated from census data) from the number of businesses with 1-9 employees.

NOTE ON THE ESTIMATION OF TECHNOLOGICAL COEFFICIENTS AS METHODOLOGY TO INTEGRATE PARTIAL RESPONSES GIVEN BY FIRMS

Stefano Pisani (ISTAT), Alessandro Viviani (University of Florence)
Alessandro Viviani Dipartimento di Statistica, Università di Firenze, viale Morgagni, 59, 50134 Firenze, Italy.

KEY WORDS: Hidden economy, partial response, reporting.

1. *Introduction*

The Italian productive structure of tradable services is characterised by a huge presence of very small firms. In order to improve the statistical basic information, some surveys have been performed about this kind of firms. One of the most difficult aspects about this survey relates to the under-reporting of the invoicing by the interviewed firms.

The phenomenon of under-reporting of the invoicing by the companies interviewed may be involved in the question of partially missing responses. This question presents, as is well known, a certain degree of autonomy with respect to that of total missing responses: the most relevant particularity is represented by the availability of information on non-respondents (Lillard et al. 1986), information that can shed light on certain aspects of the missing responses.

The hypothesis that is generally maintained in order to make up for the missing information is that the reticence of the interviewee manifests itself in a selective way, with reference to a subgroup of enquiries rather than to the questionnaire as a whole. Hence we can consider that the supplied responses constitute a non-negligible information set upon which to base the process of inferring missing data: the methods developed in this direction are justified in relation to the kind of information supplied and by the ways in which the information is to be used, tied to the need to have a unified data set, without any gaps. This stresses the empirical contents and the ad hoc nature of the most used techniques (the statistical proprieties of which still remain to be explored), all traceable back to the logic of making up for missing information, eliminating, as far as possible, the bias present in the sample data.

This paper is part of a research project, sponsored by the Italian National Statistical Institute (ISTAT), which studies the quality of the procedures for estimating the supply in the tradable service sector. In section 2 we give a general outline of the estimation procedure, and we stressed both analogies and differences with respect to the method presently followed by ISTAT. Section 3 is devoted to the illustration of various specifications used in the estimation of technological coefficients. In section 4 the methodology is applied to the under-reporting firms and we analyse the sensitivity of the results to the hypotheses underlying the technique. Concluding remarks follows.

2. *The imputation techniques*

The evaluation of the phenomenon of under-reporting brought about by ISTAT is inspired by work conducted by Franz (1985) on the underground economy estimates in Austria.

This criterion is based on the Hypothesis that the costs sustained by the firms are declared as a whole, while the partial response is considered relative to invoicing when the *per capita income of employers is less than that of the employees*. In other words, it is assumed that the costs are retrieved without systematic error for all the firms, and that the total invoiced is correctly observed only for a subgroup of the statistical units included in the sample: the criterion used for defining as partial the response obtained by the firm is based on the comparison of the per capita income received by employees and that earned by employers, under the hypothesis that is convenient to under report the total invoiced, for example to remain consistent with income tax returns, while they are not reticent regarding the explanation of costs. This procedure of correction of the total invoiced (and consequently of value added) is based, therefore, on the revision of the per capita income of the employers.

ISTAT currently applies this methodology to data derived from surveys conducted among small enterprises (defined as those that employ less than 10 workers), with the goal of estimating aggregate supply for national accounts. The revision is limited to the subgroup of smaller enterprises, insofar as it supposed that this is the size of firm on which the phenomenon of fiscal evasion and, presumably, also the under-reporting of the total invoiced are mainly concentrated.

This study proceeds from an empirical evaluation of an imputation technique alternative to that just outlined in partial responses cases.

This technique has been referred to the firms that make up the value added sample in the small business carried out in 1988. The classification of economic activity under consideration is illustrated in table 1.

Table 1. Description of economic activity sectors

Economic activity sectors	Description
723	Road transportation
831	Financial consultants
832	Insurance consultants
833	Real estate firms
834	Investments agents
835	Legal consultants
836	Accountants, tax consultants
837	Technical services
838	Advertising and public relations
839	Other business oriented services

From table 1 we can defined the sectors from 830 to 839 as Business-oriented services. The sample size is approximately 5100 firms.

In fig. 1 is illustrated the Share of under-reporting enterprises in total firms. This share is greater than 60%, and this phenomenon is mainly concentrated in that firms with less than 6 workers and in the south of Italy.

The evaluation of the phenomenon of under-reporting here analysed is based on the so called indirect methods of reconstruction of a "real" aggregate (the income) and, in a way, derives from a study carried out by Pissarides and Weber (1985) on English data. The method used is based on the hypothesis that the total invoiced is accurately observed for the sample subgroup who do not under-report (per-capita income of employers≥per-capita income of employees) and that the quantities relative to costs are observed without systematic errors in the entire unit included in the sample.

In keeping with these hypotheses, which are consistent with or at least not contradictory to those assumed by the method used by ISTAT, we can specify and estimate a relationship between the total invoiced and costs for the sample subgroup presenting complete responses.

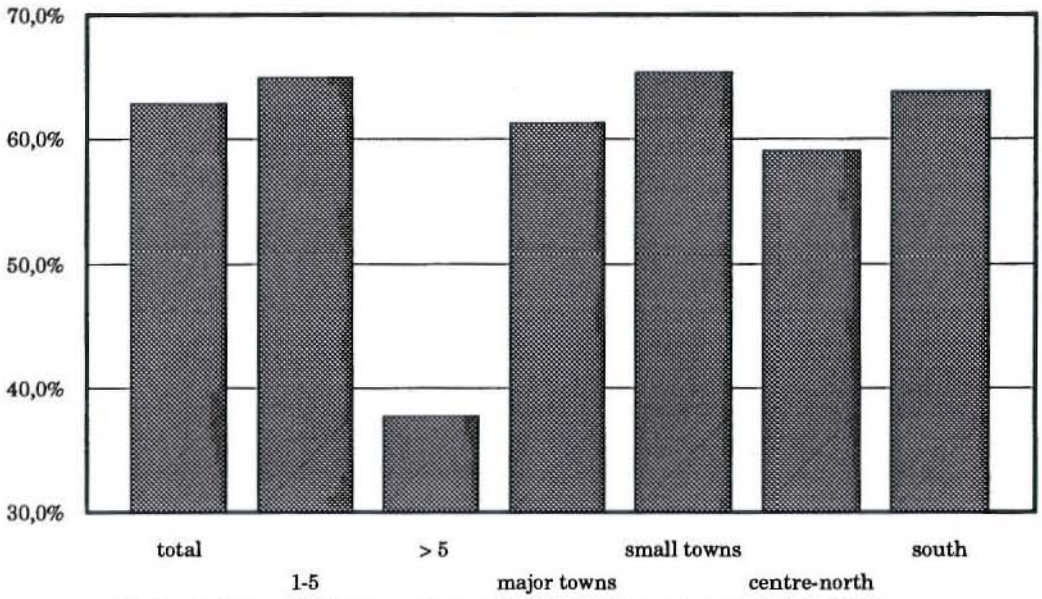


Figure 1. Share of under-reporting enterprises in total firms - year 1988

The "maintained" hypothesis is that this relationship also holds for the under-reporting enterprises, for which the relationship may be adequately used to reconstruct the "true" sum invoiced using the error-free quantities as a starting point.

As one is able to see, the analogies with the ISTAT method are easily seen in terms of the hypothesis on the data and of the process of revision performed at the desegregate level, and therefore, for each individual productive unit observed as under-reporting.

Logically speaking, the proposed method is different than that followed by ISTAT for at least two different kinds of reasons.

Above all it removes the hypothesis, considered "minimal", of revising the income of the employers on the basis of that of the employees.

Second it contemplates that the revision amount might vary on the basis of the dynamics of the gross profit margin, which may in turn be affected by various cyclical phases of the economic system as a whole as well as by different situations than can occur in the single, desegregate, productive units .

3. *The estimations of technical coefficients*

In keeping with the hypotheses presented in the previous Section, here we suggest some alternative specifications to estimate the technical production coefficients. We will then use these coefficients to compute the adjustment on the total revenue of the under-reporting firms.

As first approximation, regarding the total firms defined as non under-reporting the following equations have been estimated

$$RC=\alpha+\beta\ CV+\gamma\ AM+\delta\ IN+\zeta\ D1_i+\eta\ D2+U \quad (1)$$

Where $D1_i$ and D2 are dummy variables and i=1 or 2. In equation (1) the (RC) per-capita revenue of employers are a linear combination of (CV) (intermediate costs + labour costs) / total employment, the (AM) capital depreciation per employers and the (IN) interests per employers.

Furthermore, with the aim of more adequately representing the behaviour of the economic subject we consider the dummy variable D1 as a proxy for differences in behaviour according to the size of the enterprises. As first approximation $D1_1$ is introduced with value 0 for firms having from 1 to 5 total employment and 1 otherwise. Since this variable is seen highly significant a better approximation is sought trough the variable $D1_2$ (total employment of each firm) for the change of technical coefficients of production brought about by change in firm size.

Another dummy variable regards the geographical location of the firm. Keeping in mind the specificity of the Italian productive system, a dummy D2 is inserted with values 0 for the firms that have their headquarters in the Centre-North and 1 for those in the South and on the Islands.

The relationships are estimated with logarithmic specifications of the variables, in keeping with Theil 1971, according to which a logarithmic formulation has a lower residual variance with respect to that for other functional forms with the same dependent variables, in view of a absence of a priori knowledge on the forms of relationships studied. Even from this point of view it is necessary to underline the experimental nature of this exercise.

The estimates obtained by equation (1) are illustrated in table 2.

Table 2. Estimated technical coefficients of production for small enterprises -1988

	723	831(1)	833(2)	835(3)	838	839
α	2.499	2.827	3.340	3.076	3.091	3.174
	(42.5)	(37.5)	(41.1)	(31.9)	(23.7)	(39.4)
β	0.567	0.465	0.383	0.445	0.423	0.439
	(31.2)	(19.2)	(14.2)	(13.6)	(10.7)	(16.2)
γ	0.010	0.124	0.138	0.206	0.204	0.202
	(10.1)	(5.0)	(4.9)	(6.3)	(3.7)	(7.1)
δ	0.111	0.107	0.130	0.065	0.160	0.116
	(9.6)	(4.3)	(5.6)	(2.2)	(2.9)	(4.3)
ζ	0.515	0.401	0.064	0.252	0.320	0.652
	(30.1)	(9.4)	(7.5)	(4.7)	(3.7)	(7.8)
η	0.097	0.155	-0.401	-	-	-
	(2.5)	(2.6)	(-2.6)	-	-	-
N	928	409	246	170	87	230
R^2ad	0.912	0.713	0.774	0.813	0.817	0.819
RMSE	0.412	0.479	0.593	0.430	0.486	0.465

(1) Includes categories 831 and 832. (2) Includes categories 833 and 834. (3) Include categories 835,836 and 836. "-" indicates that the coefficient was not considered due to low level of significance. For 835 and 839 the size variable $D1_1$ was used, while for all other $D1_2$ was used. R^2ad= adjusted R^2. RMS= root mean square error. T values are reported in parentheses.

In order to improve the fitting of the theoretical model to the observed data, we adopted a transcendental logarithmic (trans-log) specification (Fuss et al. 1978, Jorgenson 1986). The trans-log function can be envisaged as a second-order Taylor's series approximation in logarithms to an arbitrary function. We adopted the trans-log specification because it is more general and flexible than the linear one (1). Furthermore, it does not impose a priori hypotheses on the scale returns. We consider the same function of equation (1) without the dummy variable D2, that can be written, in general form, as

$$RC = f(CV, AM, IN, D1)$$

where D1 is the total employment of each firm.

The trans-log, as usual, is specified as

$$\begin{aligned} RC = \alpha &+\beta_1 CV+\beta_2 AM+\beta_3 IN+\beta_4 D1+0.5(\beta_{11} CV^2+ \\ &+\beta_{22} AM^2+\beta_{33} IN^2+\beta_{44} D1^2)+\beta_{12} CV\, AM+ \\ &+\beta_{13} CV\, IN+\beta_{14} CV\, D_1+\beta_{23} AM\, IN+ \\ &+\beta_{24} AM\, D1+ \beta_{34} IN\, D1 \quad (2) \end{aligned}$$

The main results obtained from equation (2) are illustrated in table 3.

Table 3. R^2 adjusted and root mean square error derived by estimates of equation 2

Sectors	R^2ad	RMSE
723	0.945	0.325
831	0.834	0.366
833	0.850	0.418
835	0.867	0.361
838	0.895	0.366
839	0.890	0.361

On the basis of the statistics showed in table 3, we can conclude that equation (2) fits better than equation (1) the observed data.

4. *The imputations of the firms defined as under-reporting*

Two imputations of the firms defined as under-reporting were carried out using the estimates of parameters from (1) and from (2). In this way, two revisions of the revenues were obtained. On the basis of this revised item, it was possible to obtain a new estimate of value added.

Since for both equations (1) and (2) we used the logarithmic transformation for variables, to reconstruct an unbiased value for the total amount of the revenues for under-reporting firms (RC*) we needed to apply the following correction on the log of the under-reporting firms' revenue (RCS) obtained from the two equations (Harvey, 1989, page. 222).

$$RC^* = \exp(RCS+\sigma^2/2)$$

Furthermore, after having reconstructed the value added of the single economic sectors, summing the observed data for non under-reporting firms to those estimated for the under-reporting firms, we applied a filter to eliminate outliers. For that, we defined as outliers those firms showing a per-capita valued added higher than $\pm 2.5\ \sigma$ the corresponding value computed for the proper economic sector and for the class of total employment 1-5 and higher than 5.

The overall effect of the revision procedure is reported in table 4.

Table 4. Rate of Revaluation of Value Added

Sectors	equation 1	equation 2
Road transport	47.7%	47.8%
Business service	62.3%	60.4%

From the examination of table 4, one can see that the results obtained from the two equations are stable enough. Moreover, note that the size of re-evaluation for road transport is around the 50%, whereas for business service is about 60%. This re-evaluation, which is to be considered as a provisional results since further investigation is needed, is due to two main reasons: the large diffusion of under-reporting (Figure 1) and the hypothesis used to classify a firm as under-reporting.

On the basis of the latter hypothesis, discussed in section 2, we implicitly assume that small enterprises without a satisfactory revenue will stop their economic activity. Even if such a behaviour is partially confirmed by the elevated natality and mortality among the small firms producing services, we cannot exclude a priori that firms experiencing a temporary crisis will not keep on operating on the market.

In order to consider the latter aspect, we suggest a simple exercise in order to reduce the threshold on the basis of which a firm is defined as a firm under-reporting. In this first phase we proposed an ad-hoc hypotheses on the basis of which a firm is not considered as under-reporting even if the income of the employer is smaller than 10% of her employees' income. In so doing we increased by 26%, respectively, 18%, the set of the non under-reporting firms for the road transport and the business services sectors. On this newly defined population, equation (2) was re-estimated, obtaining a re-evaluation of the per-capita

value added equal to 38% for the road transport and 50% for the business services. From this results it is apparent that the size of re-evaluation is heavily dependent on the definition of non under-reporting firm. Thus, as a line of future research it would be interesting to pursue a mixed cross-section/time series approach, in order to establish a threshold separating the two sets of firms on the basis of the observed behaviour. In that, we should consider also the firms which suffer losses which may derive from management reasons or from the various phases of the economic cycle.

5. *Concluding remarks*

From the results just presented it is clear need for a check of economic consistency on the data obtained from the surveys on the small firms. In reference to the Italian case, in fact, there is the suspicion that the surveyed firms respond to statistical surveys in a way consistent with tax returns. Since the fiscal evasion phenomenon is particularly concentrated in small firms, it is plausible to assume that also the total amount invoiced resulting from the statistical surveys be undervalued.

In particular two sectors of activity were examined (road transport and business services) which are characterised by a very high number of very small firms (the share of firms with less than 10 workers in total firms is about the 80%). On this set of firms, specifically surveyed by ISTAT in 1988, a revaluation methodology, originally proposed for the Austrian economy, was applied.. Here, we suggested an alternative methodology of re-evaluation based on the estimation of technical production coefficients.

To apply this techniques, we assume first the same definition of under-reporting firms as in ISTAT method. Such a definition is based on the comparison between per-capita income of employees and the employers, when the latter is at least as large as the former. Thus the information obtained by the under-reporting firms are considered as partial and therefore in need of integration.

The first step for such an integration was the estimation of a technical coefficient function using both a linear and a trans-log specification on the set of non under-reporting firms.

We used the estimated coefficients to impute the data to the under-reporting firms obtaining a new estimate of the value added for the whole economic sector considered. The re-evaluation coefficients obtained by this exercise seem to confirm the fact that under-reporting has serious consequences as non-sampling error in building the supply-side macroeconomic aggregates.

The results presented in this paper cannot be considered as conclusive, because they require further analysis from both the theoretical and empirical points of view. From these results, it should be clear that the theoretical model adopted can prove to be useful criterion of control and validation of the data derived from surveys.

The analysis carried out in Section 4 shows how sensitive the results of the correction methodology are to the definition of an under-reporting firm. Following up on this result, it would be interesting to supplement the proposed analysis with a time-series analysis, referring also to bigger-sized firms, which allows to modify the definition of under-reporting firm on the basis of the phase of the cycle experienced by the entire economy.

Finally, as future development of this techniques of imputation, it would be useful to deepen the line of research suggested by Frey and Weck-Hanneman (1984) and Barthelemmy (1988), among others, according to whom the evaluation of under-reporting can be considered as a latent statistical variable.

References

Barthelemmy P. (1988), "The macroeconomics estimates of Hidden economy: a critical analysis", *The Review of Income and Wealth*, 34, 2.

Franz A. (1985), "Estimates of the hidden economy in Austria on the basis of official statistics", *The Review of Income and Wealth*, 4.

Frey B. S. and Weck-Hanneman H. (1986) "What do we really known about wages? The importance of non reporting and census imputation", *Journal of Political Economy*, 94.

Fuss M., Mc Fadden D. and Mundlak Y. (1978), "Survey of functional forms in the economic analysis of production", Fuss M. and Mc Fadden (eds.) *Production Economics: A dual approach to theory and applications*, vol. I, North Holland, Amsterdam.

Harvey A. C. (1989) *Forecasting structural time series models and the kalman filter*, Cambridge University Press.

Jorgenson D. W. (1986) "Econometrics method for modelling producer behaviour", Grillicers Z. and Intrilligator M. (eds.), *Handbook of Econometrics*, North Holland, Amsterdam.

Lillard L., Smith J. P. and Welch F. (1986), "What do we really known about wages? The importance of non reporting and census imputation", *Journal of Political Economy*, 94.

Pissarides C. A. and Weber G. (1989), "A expenditure-based estimates of Britain's black economy", *Journal of Public Economy*, 39.

Theil H. (1971), *Principles of Econometrics*, North Holland, New York.

LISTING FRAMES AND MAPS IN AREA SAMPLE SURVEY ON ESTABLISHMENTS AND FIRMS

Alessandra Petrucci, Monica Pratesi, Università di Firenze
Monica Pratesi, Dipartimento Statistico, Viale G.B. Morgagni, 59, Firenze, Italy

KEY WORDS: area sampling, geocoding, business registers.

1. INTRODUCTION

The question of whether there are viable alternatives to the area sample to cover businesses not represented by the list is a perennial one for many Central Bureau of Statistics [7]. The purpose of this paper is to document the potential uses of an area frame for current surveys on establishments and firms in Italy.

In section 2 of the paper, we present a method of building an area frame for business establishment surveys, using a Geographical information System (GIS) for geocoding lists of establishments derived from administrative files [1] [11]. In section 3 we report the results of some simulations of different area sampling schemes. The goal is to evaluate which of this samples scheme give the best results in the estimation of the total of the employees for municipalities.

The application is carried out using data of establishments files of the Chamber of Commerce and Industry and the list of enterprises derived from files of the National Institute of Social Security for a group of Municipalities of the province of Pistoia (Italy). Enumeration districts of the 1991 Census of population and further segmentation of them will be used as sampling areas.

2. THE AREA FRAME

This work proposes some results obtained for four Municipalities of an Italian region Agliana, Montale, Serravalle and Quarrata in the province of Pistoia (Tuscany), taken as a case study for the testing of the computerisation of the **topographic plan of the decennial census**. Such a test was thought to be useful as Italian National Institute of Statistics (ISTAT) as well is going to begin an analogous work [9] which, starting from the segment sheets, leads to the building of the **national computerised street directory**, conformed with other Countries choices [4].

Other Countries' experience and applications carried out in the past by ISTAT itself, suggest the need of segmenting the **Enumeration Districts** (EDs) in smaller areas. Looking at the census materials (maps and forms) produced by the Municipalities on the occasion of the census survey, the first micro areas to be proposed as enumeration districts segments seem to be the street segments. The term **Street Segments** (SSs) indicate the road arcs which form the enumeration district and constitute the path followed by the enumerator in delivering or retiring the questionnaire.

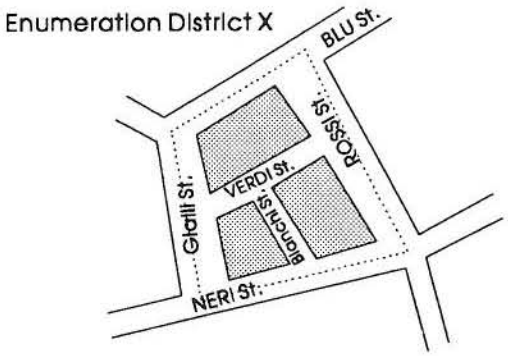
The street segments of each enumeration district are listed in the auxiliary form called **segment sheet**.

The idea is to relate the digital map of the road network of the Municipalities with the topographic plan of the census. The road network can be extracted from the regional large scale numeric cartography (1:2000), processed by the Tuscany Region [12]. The problem is that the digitised arcs have not the provincial road code attribute. To assign the provincial road code to them the street name is taken to account. To add the enumeration district code to the arcs contained in the EDs, the overlay of the digitised maps of the enumeration districts boundaries on the road network, extracted from the numeric cartography, was used. In such a way, the double coding "enumeration district code - provincial road code" is obtained for each road and this allows the exact linkage between the segment sheets and the attributes of road arcs of the digital map.

The final result is the digital map of the road network divided in SSs and EDs. The quality of results depends on the accuracy of the previously described operations. Particular attention should be reserved to the roads delimiting the enumeration district. They work as boundaries of enumeration district and they are, at the same time, SSs of the enumeration district itself (dashed line in figure 1).

For such roads, especially in the urban EDs, the road axis is the boundary between two enumeration districts and checks on the correctness of the attribution of address numbers to the right or to the left side of the street are often necessary.

Once the operations are over, any data set containing full address of the elementary units, i.e. provincial road code and address number, can be linked through the provincial road code to the road



SEGMENT SHEET
ENUMERATION DISTRICT # X

STREET CODE	STREET NAME	ADDRESS # FROM-TO	ADDRESS # FROM-TO
1500	VERDI ST.	1 9	2 10
1001	ROSSI ST.	- -	20 40
1002	NERI ST.	11 41	- -
1003	GIALLI ST.	89 103	- -
800	BLU ST.	- -	70 100
570	BIANCHI ST.	1 9	2 8

fig. 1

network, with consequent attribution of the included data to the street segments.

Thus, it is possible to stratify the SSs on the basis of any variable included in an administrative file referring to the firms.

Geocoding a list of establishments and firms means that for each unit its spatial position, referred to a geographical coordinate system, should be assigned [5]. For the case study, the available information recorded in the Geographical Information System (GIS) are schematically:

- the road network constructed as described above;
- the list of elementary units (establishments and firms) fully coded respect to the address field.

By means of the address field, each unit is assigned to the opportune SS and placed in correspondence of its address number. The procedure carrying out this operation is called address matching [14]. Once the unit is geocoded, also all the referenced information included in the file are automatically associated with its geographical location. The result is then the building of maps of the spatial distribution of the filed variables. Two list of frames have been geocoded: the list of Chamber of Commerce and Industry (CC) and the list of the National Institute of Social Security (NISP). The CC list is one of the best to identify establishments from a statistical point of view. It is not a customer file, all the active establishments are the province can be faced, the classification of the economical activities is that officially used by ISTAT. The NISP list has not the complete coverage of the target population, because contains only the employers firms and establishments. Such files are often incomplete and sometimes include data of uncertain quality but are anyway a useful reference, quicker than the census survey for upgrading of layering of the enumeration districts and of the street segments they could be decomposed into [8].

All the operations described above were carried out by mean of the GIS Arc/Info [10].

3. RESULTS

The comparison between area sampling and dual list sampling has not been done completely up to now. Here we report only the results of the simulations of area sampling schemes and some preliminary results of the construction of dual frame (list frame and area frame) for one Municipality (Agliana).

In the Municipalities under study (Agliana, Montale, Serravalle and Quarrata) small firms, in term of employees, working in textile and precision

Table 1.

Municipality	# EDs	# EDs with establishments	# SSs with establishments	# SSs digitised
Agliana	193	176	432	381(1)
Montale	54	47	221	-
Quarrata	118	109	490	354(2)
Serravalle	66	58	193	-

Notes:
(1) SSs for those the procedure of linkage with the segment sheet turned out well.
(2) Type 1, from which digitised 210 (see note 1).
From: our elaboration

Table 2.

Municipality	Establishments		Employees	
	Census	C. Commerce	Census	C. Commerce
Agliana	1505	1700	4988	5139
Montale	1047	1129	3263	3418
Quarrata	2859	3103	3007	3231
Serravalle	667	821	8852	9262

From: ISTAT - VII Census of Industry Trade and Service - October 1991 - Provisional data and Chamber of Commerce of Pistoia file at 31/10/91

mechanical industries are prevalent. The VII Census of Industry Trade and Service counted 6078 active establishments with 20110 employees.

In the Municipalities are active the 24.68% of the establishments of the whole province, with the 21.23% of the employees of the province. More than 90% of the establishments are small (1-10 employees). The distribution of the employees is concentrated in textile industry (70%) and in the sector of commerce (20%). The Municipalities are characterised by the prevalence (90%) of small businesses with high birth (death) rates of employers and non employers establishments.

Area Sampling Results

To investigate the performance of estimate of total employees and of the average of the employees for establishments, in area sampling, some simulations have been carried out to evaluate the efficiency of the estimators. The choice of probability selection is adequate to the spatial distribution of small business in the case study [2] [3].

Enumeration districts have been considered as primary sampling units (PSUs) and street segments as PSUs as well. The comparison between different designs are done for fixed sampling fraction (f) and for fixed expected number of elementary units in the sample ($E[n_S]$). The elementary units, establishments, are considered as secondary stage units (SSUs). The following designs have been simulated:

1. two stages sampling with equal probability selection and PSUs = EDs or PSUs = SSs;
2. two stages sampling with variable probability selection and PSUs = EDs;
3. proportionate stratified two stage sampling with the following stratification criteria of the PSUs:
   a. EDs type;
   b. prevalence of economic activity in EDs
   c. prevalence of legal form;
   d. prevalence of legal form and economic activity.

Design **1**: street segments are better PSUs than EDs for the same number of elementary units in the sample. This is due probably to the fact that the number of PSUs in the sample is bigger when we use smaller PSUs.

Design **2**: the relative variances of estimator of average resulting from some simulations are show in table 5 and table 6.

For Agliana Municipality, the selection of PSUs = EDs with probability proportional to 1/d (where d is the Euclidean distance between the EDs and the mean center of the municipality) gives good results (V'(**2**)) [13].

Design **3**: the criterion of stratification, with PSUs = EDs, which gives more interesting results is *a*, the efficiency of the estimator becomes better if we complicate the design **3** with selection with probability proportional to 1/d.

We have implemented it for the Municipality of Quarrata. In this case, the PSUs = SSs have been

Table 3.

Municipality	Industry	Business	Other Activities	Institution	Total
Agliana	907	305	249	44	1505
Montale	580	255	161	51	1047
Quarrata	1512	764	494	89	2859
Serravalle	314	211	110	32	667
Pistoia	8017	8943	6258	1414	24632

From: ISTAT - VII Census of Industry Trade and Service - October 1991 - Provisional data

Table 4.

Municipality	Industry	Business	Other Activities	Institution	Total
Agliana	3136	750	551	551	4988
Montale	2062	565	355	281	3263
Quarrata	5395	1640	1134	683	8852
Serravalle	1903	720	193	191	3007
Pistoia	37526	23550	18410	15424	94730

From: ISTAT - VII Census of Industry Trade and Service - October 1991 - Provisional data

stratified by type of EDs. Different selection criteria have been used in each stratum:

- in urban stratum, two stage sampling with PSUs = SSs selected with probability proportional to 1/d has been simulated;
- in *nucleo abitato*[1] stratum, two stage sampling with PSUs = SSs selected with equal probability has been simulated;
- in *case sparse*[2] stratum, one stage simple random sampling of SSs has been simulated.

The relative variances are lower, at the same expected sample size, than those obtained in design with stratification criterion (**3.***a*).

Dual Frame Construction

Relying only on NISP list for the Municipality of our case study can produce substantial coverage bias due to the fact that the list tends to miss many of the new business especially the smallest business which are likely to be non employers establishments. The idea is to construct a design that integrates the NISP list with an area sample to supplement the list.

To simulate the dual frame sampling design the first step is geocoding NISP list. Direct geocoding is difficult because the provincial road code is not a field of the NISP file. Geocoding is possible trough the linkage of NISP file with the CC file. For the Municipality of Agliana the NISP register at December 31, 1990 counts 385 establishments (385 unilocalized firms), the CC file at the same data counts 1525 establishments. The results obtained are shown in table 7.

At present the dual sampling study is in progress.

4. CONCLUSIONS

The work indicates that the methodology used to build an area frame from census materials is applicable to other Municipalities as well. The methods used are satisfactory both in forming street segments and in doing this for a fairly low effort.

Although the development of an area frame is not cheap, it is not as expensive as it once was since one can use jointly the national computerised directory and the census maps to outline the boundaries which would reduce the resources required to prepare maps.

In area sampling data are collected from businesses within the selected areas. Since one select clusters of businesses in the sample to be surveyed, an area frame have the inefficiency attributed to cluster sampling [6].

However, using data from businesses registers list frame and census maps to stratify the areas and calculating adequate probability of selections, can reduce the inefficiency at least partially.

---

[1] *nucleo abitato* = a group of neighborhood houses with at least five families and which the distance between each other is not greater than 30 meters.

[2] *case sparse* = country-houses or houses located at long distance between each other.

Table 5. Municipality of Agliana.

		V'(2)	V'(3.*a*)		V'(3.*d*)	
f	E[n $_S$]	EDs	EDs	SSs	EDs	SSs
0.20	340	0.015	6.90	13.12	5.03	10.18
0.30	510	0.010	6.08	11.51	4.43	8.90
0.40	680	0.001	5.20	9.80	3.78	7.57

Table 6. Municipality of Quarrata

		V'(2)	V'(3.*a*)		V'(3.*d*)		V'(3.*a*)
f	E[n $_S$]	EDs	EDs	SSs	EDs	SSs	EDs ,SSs
0.20	620	0.003	6.62	18.17	9.33	17.17	3.02
0.30	930	0.003	5.79	15.85	8.16	14.97	2.13
0.40	1240	0.001	4.96	13.56	6.99	12.80	1.98

Table 7.

Key-item	Matched
V.A.T.	152
Fiscal Code	142
Residential Address	90
Not possible to locate	1
Total	385

Even if we have not completed our work yet, it is felt that an area frame properly designed to supplement a list frame would improve the chances of producing better estimates.

Acknowledgement

This work has been supported by a research contract (n. 93 12/09/87) between the National Statistical Institute (ISTAT) and the Department of Statistics of the University of Firenze (Italy).

REFERENCES

[1]ARBIA, G. (1991). GIS-based sampling design procedures. *Proceedings of the Second European Conference on Geographical Information Systems*, Brussels, April, 2-5 1991, pp. 27-35.

[2]BIFFIGNANDI, S. (1986). Modelli di organizzazione spaziale dell' industria. *Atti SIS*, Bari, pp. 215-227. (in Italian)

[3]BRACALENTE, B. (1990). Il censimento dell' industria e le indagini economiche correnti in Italia. *Atti SIS*, Padova, pp. 179-189. (in Italian)

[4]CERTOMÀ, C.A. (1992). L'utilizzazione del telerilevamento da parte dell' ISTAT: realizzazioni ed orientamenti. *Seminario su "L'impatto del telerilevamento sul Sistema Statistico Europeo"*, Bad Neuenahr (Germany), Settembre, 22-24 1992. (in Italian)

[5]GOODCHILD, M. (1984). Geocoding and Geosampling. In *Spatial Statistics and Models*, G.L. Gaile and C.J. Willmott (eds), D. Reidel, Dordrecht.

[6]KISH, L. (1965). *Survey Sampling*, Wiley, New York.

[7]KONSCHNIK, C.A. and KING, C.S. (1992). *Reassessement of the use of an area sampling for the Monthly Retail Survey*, Bureau of Census, Washington DC..

[8]MARTINI, M. e AIMETTI P. (1989). *Un archivio delle imprese per l' analisi economica*, Uniore Regionale delle CCIAA della Lombardia, Milano. (in Italian)

[9]ORASI, A. e GARGANO, O. (1992). Le basi territoriali dei censimenti 1991: il progetto CENSUS, *Atti AM/FM-1992 Italia*, Firenze, Novembre, 16-18 1992, pp.38-57. (in Italian)

[10]PETRUCCI, A. e PRATESI, M. (1992). *Liste di imprese: loro riferimento territoriale per campioni areali ed altre analisi spaziali*, Working Paper n. 40, Dipartimento Statistico, Università di Firenze, Firenze. (in Italian)

[11]SAARFELD, A. (1991). Construction of spatially articulated list of frames for household surveys, *Proceedings of Symposium 91 - Spatial Issues in Statistics*, Ottawa, November, 12-14 1991, Statistics Canada.

[12]PELACANI, G. (1990). *Tavola dei contenuti, segni grafici e codici per cartografia a scala 1/2000, versione 2.1*, Regione Toscana - Giunta Regionale, Dipartimento Urbanistica - Servizio Cartografico, Firenze. (in Italian)
[13]SÄRNDAL, C.E., SWENSSON, B. and WRETMAN J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.
[14]SHOLTEN, H.J. and STILLWELL, C.H. (1990). *Geographical Information Systems for Urban and Regional Planning*. Kluwer, Dordrecht.

INTRACLUSTER RATE OF HOMOGENEITY AND DIMENSION OF AREAS IN AREA SAMPLING

Corrado Lagazio, Monica Pratesi, University of Florence
Monica Pratesi, Dipartimento Statistico, Viale Morgagni, 59, 50134 Florence, Italy

KEY WORDS: area sampling, cell size, homogeneity.

SUMMARY. The variance of the estimator of a total in single stage cluster sampling depends on the homogeneity of the study variable in the PSU's. A similar connection can be found in area sampling, where clusters are cells (quadrats) of equal extension, containing a different number of elementary units. In this paper, the relation between the intra-areal rate of homogeneity and the size of the cells is studied under a superpopulation model in order to minimize the variance of the estimator of a total.

1. Introduction

In area sampling of business establishments, a crucial point is the size of areas both in terms of their extension and of the number of elementary units they contain, because the variance of the estimator of a total depends on the variance within the cells. When the study area can be divided in subareas by the researcher (e.g. through the superimposition of a grid of quadrats), a problem to be faced with is cell extension, in order to reach lower levels of the variance of the estimator of the overall total.

In a model assisted approach, the distribution of elementary units in the study area can be described by probabilistic models of spatial distribution. In the paper, the spatial locations of business establishments in a study area is assumed to follow a Negative Binomial model (concentrated pattern). Further, a superpopulation model, conditional on distribution of the elementary units, is assumed for the study variable. On the basis of the proposed models, the relationship between the intra-areal rate of homogeneity and size of areas is studied and some preliminary results are presented.

2. The intracluster rate of homogeneity

Let $Y = (Y_{ij})$ be a quantitative variable associated with N elementary units geographically distributed on an area A. The area is divided into $J(\alpha)$ cells (clusters) of equal size α, the j-th containing N_j elementary units, so $i = 1, 2, \ldots, N_j$ indexes the elementary units of the j-th cell and $j = 1, 2, \ldots, J(\alpha)$ indicates cell membership.

Let SST be the deviance of Y, SSW the deviance within clusters and SSB the deviance between clusters. The Standard ANOVA decomposition gives us the following result:

$$SST = SSW + SSB$$

where

$$SST = (N-1)\sigma^2$$

$$SSW = [N - J(\alpha)] \sum_j \sigma_j^2$$

σ^2 is the overall variance of Y and σ_j^2 is the variance of Y in the j-th cluster. Since $SSB \geq 0$, it follows that

$$\frac{\sum_j \sigma_j^2}{\sigma_2} \leq \frac{(N-1)}{[N - J(\alpha)]}$$

In the present context we define the rate of homogeneity δ as

$$\delta = 1 - \frac{\sum_j \sigma_j^2}{\sigma_2}$$

We know (Särndal et al. (1992), Hansen et al. (1953)) that in simple random sampling of clusters the *deff* of $\hat{t}_\pi$, estimator of the overall total of the study variable, can be expressed as follows

$$deff = 1 + l_1\delta + l_2$$

where

$$l_1 = \frac{N-1}{N - J(\alpha)}$$

and

$$l_2 = \frac{cov(N_j, \sum_i y_{ij}^2/N_j)}{[N/J(\alpha)]\sigma^2}$$

3. The superpopulation models

We will assume that the number of business establishments in the j-th cell follows a negative binomial distribution with parameters K and P, namely

$$\Pr[N_j] = \binom{K+n-1}{K-1} (P/Q)^n (1 - P/Q)^K$$

where $Q = 1 + P$. This model is frequently used to describe concentrated patterns resulting both from real and apparent contagion processes (Rogers 1974, Gori 1982). Suppose to partition the study area into

quadrats and to fit a negative binomial with parameters K_1 and P_1 to the frequency of items in quadrats. Cliff and Ord (1981) have developed an approximate procedure to distinguish between these two kinds of processes. Under suitable assumptions, they have found that, if blocks of s original cells are combined to form new larger quadrats, the resulting frequency distribution of items in the new cells is still negative binomial, but with parameters K_s and P_s. If the original negative binomial has been generated by a true contagion process, then the relation between the parameters in the original and new cells is

$$K_s = sK_1$$
$$P_s = P_1$$

while, in a spurious contagion situation, the relation is

$$K_s = K_1$$
$$P_s = sP_1$$

These are only approximate descriptions of the connection between the parameters of negative binomial and the cell size (Hodder and Orton 1976). To be consistent with Cliff and Ord (1981) procedure, the expected value of the number of business establishments is chosen to be directly proportional to the area of the cell with respect to the total area. Then, in what follows, we assume that the expected value of the number of in the j-th cell is

$$E[N_j] = KP = \frac{\alpha}{A}c \tag{1}$$

where c is a superpopulation parameter which can be seen as the expected value of the total number of business establishments in the whole study area.

Let y_{ij} be the number of employees in the i-th business establishment of the j-th cell. We assume that, conditional on N_j, y_{ij} is Poisson distributed with conditional mean θ_j, where θ_j is equal to

$$E[y_{ij}|N_j] = \theta_j \begin{cases} \frac{\alpha}{AN_j}\theta & N_j \neq 0 \\ 0 & N_j = 0 \end{cases}$$

In this way we suppose that the expected number of employees of the i-th business establishment is inversely proportional to the density of business establishments (defined as N_j/α) in the j-th cell.

4. Results

Under the hypotheses above specified, after some algebra, the expected value of SSW (the deviance within the clusters) can be expressed as follows:

$$E[SSW] = \theta\frac{\alpha}{A}\sum_{j=1}^{J(\alpha)}\left[1 - p(0) - \bar{E}_{\underline{N}}\left(\frac{1}{N_j}\right)\right] \tag{2}$$

where:

$$p(0) = \Pr[N_j = 0]$$

and

$$\bar{E}_{\underline{N}}(\cdot) = \sum_{r=1}^{\infty} r\Pr[N_j = r]$$

This term can be calculated using the approximation proposed by Govindarajulu (1962) (see also Johnson and Kotz, (1969), pag 136). Substituting these expressions in equation (2) we have:

$$E[SSW] = \theta\left\{\left[1 - (1+P)^{-K}\right] - \frac{\left[1 - (1+P)^{-K}\right]}{KP - (1+P)}\right\} \tag{3}$$

In the same way, we can derive the expected value of SST:

$$E[SST] = \theta\frac{\alpha}{A}\sum_{j=1}^{J(\alpha)}\left[1 - p(0) - (1+\theta)\bar{E}_{\underline{N}}\left(\frac{1}{\sum_j N_j}\right) + \frac{\alpha\theta}{A}\bar{E}_{\underline{N}}\left(\frac{1}{N_j}\right)\right] =$$

$$= \theta\left\{\left[1 - (1+P)^{-K}\right] + \frac{\alpha\theta}{A}\frac{1 - (1+P)^{-K}}{KP - (1+P)} - (1+\theta)\frac{1 - (1+P)^{-\frac{A}{\alpha}K}}{\frac{A}{\alpha}KP - (1+P)}\right\} \tag{4}$$

The expected value of SSW and SST depends on the superpopulation parameters c and θ; and on α/A, the relative size of the cells of the lattice, both directly and by K and P (see equation (1). Under the specified superpopulation model, $E[SSW]$ and $E[SST]$ can then be used to study the behaviour of the *deff* as a function of the rate of homogeneity. Defining

$$\delta^* = 1 - \frac{E[SSW]}{E[SST]} \tag{5}$$

we have a tool to give an approximated description of the intra-areal homogeneity when the parameters of the superpopulation model and α/A vary.

5. Conclusions

Substituting equations (1), (3) and (4) in equation (5) we can study the behaviour of δ^* when α/A and θ change, for fixed c. We see that, under the assumed model, the coefficient δ^* is a monotonically decreasing function of the relative size of cells (fig. 1). This result is still valid for different values of

the parameter θ and is coherent with our knowledge from sampling theory.
In *deff* formulation also l_1 and l_2 play an important role in determining *deff* values when K, P and α/A change. The problem is to evaluate the trade-off between l_1, l_2 and δ^*, looking for a relative cell size α/A that gives lower levels of the variance of $\hat{t}_\pi$ in our case. This part of the work is still in progress.

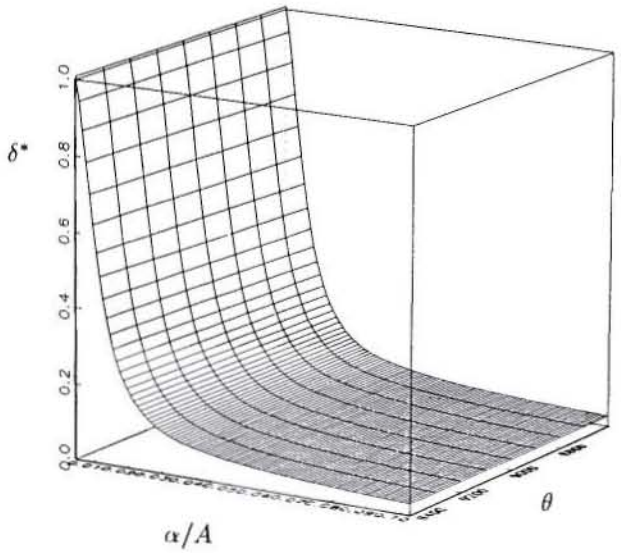


Figure 1

References

CLIFF A.D., ORD J.K. (1981), *Spatial Processes Models and Applications*, Pion, London.

GORI E. (1982), *Modelli Stocastici per l'Analisi Spaziale*, Serie Ricerche Teoriche n. 5, Dipartimento Statistico, University of Florence.

GOVINDARAJULU Z. (1962), The Reciprocal of the Decapitated Negative Binomial Variable, *Journal of the American Statistical Association*, **57**, 906-913.

HANSEN M.H., HURVITZ W.N., MADOW W.J. (1953), *Sample Survey Methods and Theory*, vol I and II, Wiley, New York.

HODDER I., ORTON C. (1976), *Spatial Analysis in Archeology*, Cambridge University Press, Cambridge.

JOHNSON N.L., KOTZ S. (1969), *Discrete Distributions*, Wiley, New York.

ROGERS A. (1974), *Statistical Analysis of Spatial Dispersion*, Pion, London.

SÄRNDAL C.E., SWENSSON B., WRETMAN, J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.