Sampling, Maintenance, and Weighting Schemes for Longitudinal Surveys: a Case Study of the Textile and Clothing Industry

by

Michele Lalla

March 1999
Abstract: A longitudinal survey of textile and clothing firms in Emilia-Romagna (Italy) was conducted using a "replenished" panel survey to produce reliable cross-sectional estimates. The firms (sampling units) were stratified according to size and age, and in all phases, dead and/or emigrated firms were considered as having already left the sample, while new units selected from newborn and immigrant firms, were considered as having already entered the sample. Population dynamics change firm composition over time and the methods of cross-sectional estimation must be adapted to the longitudinal features of the samples. This paper reviews some weighting schemes used in pursuing the aims of repeated sample surveys and presents application trials conducted on real and simulated data.

Key words: Longitudinal survey, Cross-sectional estimate, Replenished panel survey, Stratified sampling, Nonresponse, Gross change, Textile and clothing industry.

Acknowledgements: The author would like to thank the R&I firm (which conducted the survey) for supplying the data and Monica Baracchi (a researcher at R&I) for her useful information on data collection and file structure. Special appreciation is due to Sebastiano Brusco for his helpful contribution concerning the sampling design and the aims of the survey. I alone am responsible for any errors. Partial financial support was provided by the Italian MURST grant 40% (directed by Ugo Trivellato).
1. Introduction

The successive selection of samples (repeated surveys) is often used to follow up on changes in the characteristics of a population over time (Duncan and Kalton, 1987). For example, this may concern a population of establishments for which there is interest in following the trends in product destination. An opposite strategy, named panel survey or longitudinal survey, concerns the use of the same sample at different points of time. When panel surveys were introduced (Lazarsfeld and Fiske, 1938; Lazarsfeld, 1940), the power of collected data was immediately recognised, but only recently have they become increasingly utilised in economic and social studies (Duncan and Kalton, 1987; Kasprzyk et al., 1989; Kalton and Citro, 1993). The potentiality of longitudinal data, such as the measurement of gross change and other components of individual change, could provide a deeper understanding of the evolution of a population over time. However, the ideal strategy implies holding the sample size constant, but pursuing this objective is almost impossible due to demographic movements and it is also not convenient because the burden and the cost of the survey become considerable. Furthermore, the population changes in size and composition because new elements enter the population (e.g. newborn and immigrant firms) and existent elements leave the population (e.g. dead and emigrant firms). Therefore, any sampling strategy should take into account population changes over time to provide a likely representation of the trends of the characteristics and more reliable estimates.

A longitudinal survey should deal with three main topics: the selection of the initial sample, the maintenance of the sample, and the weighting scheme. However, the first and the third topics are common also to other types of surveys, although in a different way. The first and the second topics involve the methods for selecting units which could be subjected to some constraints (Cotton and Hesse, 1997). Firstly, sampling units should regularly be selected from births and take deaths into account. Secondly, the characteristics of units involved in sampling design and changes over time, such as the size or primary business activity, become "incorrect" and/or less and less correlated with other variables. Therefore, the estimates of the population parameters become more and more unreliable, as their variance increases progressively. To avoid this drawback it is possible: (1) to make every endeavour to conserve the current sampling units to keep the maximum number of the original units selected, (2) to select new elements from the population at each occasion after updating and calculating new probability of inclusion to account for births and deaths. However, in practice, any procedure will fatally encounter difficulties and failures because of changes in the probability of inclusion, as the composition of the strata will change, even if the inclusion probability has remained constant. Thirdly, it is often necessary to limit the response burden of the survey to a few occasions. Generally, after a fixed number of occasions a unit is replaced by another unit chosen from those elements which have still not been included in the sample (rotation over time), but here it is ignored.

Focusing our attention on sampling, maintenance, and some weighting schemes, this paper describes a longitudinal survey of the textile and clothing industry which was promoted by the Observatory of the labour market in the Region of Emilia-Romagna (Italy), to ascertain the feasibility of a periodic survey of structural developments, i.e., the changes and transformations occurring over time in the textile and clothing industry. The purpose of the survey was to provide a solid body of
knowledge for the formulation of intervention in terms of professional training and industrial policies on the basis of the evolution of the target population and the individual behaviour of the firms over time. Therefore, the principal data collected by the survey consisted of turnover, number of employees, type of firm (final firms and subcontractors), type of product (knitwear, outerwear and underwear), main market band within which the firms operate, type of final consumer (menswear, womenswear and children’s clothing), type of customer, destination of the products (domestic market and exports) and so on (see Brusco and Bigarelli, 1993).

To pursue the aims of the survey, the sampling design was based on a panel survey which was modified to carry out both longitudinal and cross-sectional analyses. A panel survey could provide a deeper understanding of the evolution of the textile and clothing industry over time. However, the sampling strategy was adapted in this case to suit the specific objectives of the study and the specificity of the industry. The survey on the first occasion was designed to obtain satisfactory estimates of the target population’s parameters. The stratification variables were the size and the age of the firms. The sampling procedure on the subsequent occasions updated the sample to account essentially for births and it appeared to be simpler than many others known in the literature (e.g., Hidiroglou, Choudhry, and Lavallée, 1991; Hidiroglou and Srinath, 1993; Armstrong, Block, Srinath, 1993; Cotton and Hesse 1997).

To obtain cross-sectional estimates of descriptive parameters for the population surveyed, it is necessary to weight the data collected. Aiming to compensate for non-coverage, for different probabilities of selection for sampled elements and for unit nonresponse, and to improve the precision of the sample estimates, weights are used essentially to conform the weighted sample distributions of certain variables to the corresponding known population distributions (Kish, 1992). However, their use is subject to heated debate when it comes to the construction of analytic models describing causal systems (Smith, 1984; Kalton, 1989; Pfeffermann, 1993; among others). On the basis of these aims, weights are defined according to a step-by-step procedure. First, a design weight is determined for each sampled element as the inverse of the probability with which it was selected. Second, the nonresponse weight is introduced to reduce the effect of nonresponse bias and the design weights of the responding sampling units are multiplied by the inverse proportion to the response rate. Third, the distribution adjusting weight is applied for the purpose of conforming the distribution of given variables in the sample (obtained after having applied the design and nonresponse weights) to the same distribution extracted from other more reliable external sources.

The draft of the plan and some results of the survey on the first occasion are illustrated in the Section 2. Some aspects of the sampling procedure for this application, the additional criteria on which the panel survey was based, and some results and weighting problems relative to the second occasion, are described in Section 3. The procedures used to obtain the cross-sectional estimates of the population total and some of its features (number of employees and turnover) are described in Section 4, in addition to the longitudinal estimates. Lastly, some concluding comments follow in Section 5.
2. The survey design on the first occasion

All decisions concerning the design of a cross-sectional survey were aimed at con­ 
structing the panel sample on the first occasion, taking into account the goals at 
hand, the resources and time available, as well as the information contained in the 
frame. The outline of the survey in terms of key concepts is as follows. Population: 
All textile and clothing firms in Emilia-Romagna. Domain of interest: Size and 
age groups. Furthermore, each firm was described at distinct time points as (1) be­ 
longing to the population, (2) exited from the population owing to closure, merger, 
bankruptcy and so on, (3) misclassified because its actual economic activity differed 
from the kind of activities of interest to the survey. Population characteristics of 
interest: Number of employees, turnover, type of firm, product, and so on. Sample: 
The units (firms) were selected from the frame provided by CERVED (the national 
network and data processing society of the Chamber of Commerce). Observation: 
Each firm in the sample received the questionnaire. However, most interviews were 
conducted by telephone by R&I of Carpi.

2.1. Description of the frame

The CERVED Register of Firms is a large and exhaustive sampling frame that lists 
all Italian firms because they are obligated to register with the Italian Chamber 
of Commerce. For each firm, this frame provides information on variables such as 
address, date of birth, number of employees, code of economic activity based on the 
NACE-related classification provided by the Italian National Institute of Statistics 
(ISTAT, 1981). NACE is the acronym denoting the general industrial classification 
of economic activities within the European Community.

The statistical unit was the firm (or enterprise), i.e., a legal organisational entity 
carrying out an economic activity that is industrially homogeneous, with or without 
independent control of its activity, and which provides a complete set of financial 
accounts. Approximately, it corresponds to the first, second, and third levels of the 
definition of an operating structure of business adopted by the European Community 
(see Council Regulation No. 696/93 of March 15, 1993). Each statistical unit was 
one (and only one) individual listing in the CERVED Register of Firms. The frame 
was constructed by selecting from the latter only the most important activities of 
the textile and clothing industry in the Region of Emilia-Romagna at the 3-digit 
and 4-digit levels of the ISTAT classification, such as the production of knit fabrics, 
production of other knit wear, the manufacture of outerwear (for men and boys; for 
women, girls and infants; not classified elsewhere) the manufacture of underwear 
(for men and boys; for women, and corsetry), and, finally, other activities related to 
the clothing industry.

Given that all the firms are registered with CERVED, frame imperfections in­ 
volve the accuracy of the information provided which includes coding errors, trans­ 
scription errors, errors introduced by or not corrected by editing. The latter pro­ 
duce untraceable firms, undercoverage and overcoverage. Undercoverage involves 
the omission of elements because of misclassification of activity. Presumably, this is 
problematic for large firms because they must declare their prevailing activity which, 
in many cases, is difficult to define. In fact, in the fifth size-class (firms with 50 em­ 
ployees and over), whose firms were all surveyed, 5 units were included that were not 
present in the frame because they had been classified as holding companies on the 
basis of their financial and non-productive activity. Overcoverage involves elements
that do not actually belong in the target population because they recently went out of business or their activity, although related in some way to the textile and clothing industry, does not fall within the objectives of the study (again, a misclassification).

Lastly, CERVED updates its records with some delay and firms sometimes fail to notify CERVED of variations, so that some data may not be reliable. However, it was not possible to come up with a perfect sampling frame due to excessive costs. Therefore, although the frame supplied by CERVED shows some imperfections because many variables are not checked or updated (Martini and Aimetti, 1989; Martini, 1990), it was used because it is the only one that includes all firms, even those without employees which were also an object of study. As a result, the frame includes about 10% of firms that did not declare the number of employees. On account of these missing data, the stratified sampling procedure adopted combined proportional and optimal allocation.

The advantages of the CERVED Register of Firms in monitoring firms over time were: (1) the facility of linking the firms between the subsequent occasions because there was a unique identifier for each one and the costs of maintaining the link should decrease considerably; (2) a firm may be a cluster of establishments (the production plants manufacturing homogeneous goods and/or services) involving a maximum overlap of samples between two occasions, which is sufficient to provide efficient estimates of net changes at low costs. However, there were also some difficulties. A firm may drastically change its type of activity without changing its name and/or its identifier, and vice-versa, it may change its name and/or its identifier without changing its type of activity. A firm may locate one or more establishments in one or more regions other than Emilia-Romagna, and vice-versa, it may locate one or more establishments in Emilia-Romagna remaining in the other region. The update of the national NACE-related classification may radically change the composition of the target population, as occurred in the mid 1990s, and generally the revised classification of economic activities cannot be fully converted to the previous one. Splits and mergers create difficulties in longitudinal analyses, while at the establishment level, these changes occur less frequently (see, inter alia, Baldwin, Dupuy, and Penner, 1992; Lavallée, 1994).

2.2. Stratified Sampling

Although the number of employees in the firms, \( Y \), changes over time, it is an ideal variate for the stratification of firms because \( Y \) itself has to be measured in the survey (Cochran, 1977, p. 101). Furthermore, the probability of a firm leaving the population might depend on its size and thus it is possible to evaluate the stability of the firms in terms of the number of employees. \( Y \) ranges from 0 to 3 in the first class interval, 4 to 9 in the second, 10 to 19 in the third, 20 to 49 in the fourth, 50 and over in the fifth. Finally, firms with missing data were included in the sixth class interval (\( I = 6 \)).

The duration of life for the firms represents a characteristic object of interest to analyse their survival and the dynamics of the population. Therefore, the age of the firm is another ideal variate for stratification, which also facilitated the survey design over time. The age groups were: 0 to 1 year, 1 to 2 years, 2 to 5 years, 5 to 10 years, 10 years and over (\( J = 5 \)). The sizes of the first two groups were chosen to include firms of brief duration or performing temporary activities and thus unable to stabilise themselves on the market (Solinas, 1995).
The number of firms, \( n_{ijl} \), that had to be selected in stratum \( ij \) (determined by the \( i \)-th size class and the \( j \)-th age group) at time \( t = 1 \), was calculated using proportional allocation, \( n_{p;ijl} \), because the number of employees was not available for strata determined by the sixth size-class, \( \{6j : j = 1, \ldots, 5\} \). To avoid a loss of precision in the strata with large firms, \( \sum_{i=1}^{5} \sum_{j=1}^{5} n_{p;ijl} \) was reassigned to the 25 strata for which the number of employees was available using the Neyman (optimum) allocation, \( n_{o;ijl} \), because the cost per unit was assumed to be the same in all the strata (Cochran, 1977):

\[
  n_{o;ijl} = \frac{\hat{n}_{1} p_{ijl} S_{ijl}}{\sum_{i=1}^{I-1} \sum_{j=1}^{J} p_{ijl} S_{ijl}}, \quad i = 1, \ldots, I - 1, \quad j = 1, \ldots, J.
\]

The subscripts \( i \) and \( j \) denote the stratum and \( l \) stands for \( t = 1 \). Moreover, \( p_{ijl} = N_{ijl}/N_{1} \) is the weight (\( N_{ijl} \) and \( N_{1} \) are the number of elements of the population in stratum \( ij \) and for all strata at \( t = 1 \), respectively), \( S_{ijl} \) is the standard deviation of the number of employees, \( I - 1 \) and \( J \) are the number of rows and columns of the two criteria of stratification (size of firm and age group, respectively). Given that \( n_{o;ijl} \) is proportional to \( S_{ijl} \), optimum allocation involves a higher fraction of units in the strata in which the variability of the number of employees is greater than the variability in the other strata. However, in the fifth size-class interval, the original allocation was greater than that of the population: \( n_{o;5j1} \geq N_{5j1} \), for \( j = 1, \ldots, 5 \). Therefore, the firms of sizes equal to or greater than 50 employees were all surveyed and the revised optimum allocation was applied only to classes 1-4 using the remaining sample size \( \hat{n}_{1} = n_{1} - \sum_{j=1}^{5} (N_{5j1} + n_{p;6j1}) \), where \( n_{1} \) is the total sample size.

Taking into account the amount of money budgeted for the data collection, \( n_{1} \) was determined assuming the relative error \( r = 0.18 \) which gave \( n_{1} = 823 \) (Cochran, 1977, p. 77). This sample size was increased because the nonresponse rate was presumed to be about 30% of \( n_{1} \) (Goyder, 1987; Hox and De Leeuw, 1994) and the number of firms in the sample was 1202, of which 41 firms could not be traced. The number of firms contacted by interviewers and the number of firms surveyed are reported in Table 1. It should be noted that only 78 of the 87 large firms in the fifth size-class interval were interviewed because 1 firm had left, 6 firms did not belong to the target population, and 2 firms did not cooperate, but no information was gathered using their balances because they were private partnerships and were not obliged to publish them.

### 2.3. Weighting for the first occasion

A self-weighting sample possesses considerable advantages consisting in reduced variances, simplicity, and robustness (Kish, 1992), but a sample drawn on the basis of the optimum allocation generally yields a smaller variance for the estimated mean or total than a sample drawn by proportional allocation. However, the sample size in each stratum was determined by either proportional or optimum allocation. Therefore, the expansion factors were easily obtained by

\[
  w_{ij1} = \frac{1}{\pi_{ij1}} = \frac{N_{ij1}}{n_{ij1}}
\]

where \( \pi_{ij1} \) is the probability of the sampled elements being selected in stratum \( ij \) at time \( t = 1 \) and \( n_{ij1} \) is the number of firms selected for the sample in that stratum at the same time.
Table 1 - Number of firms contacted by interviewers, $n_{c;ij1}$, and number of firms
surveyed or respondent, $n_{r;ij1}$, by size and age classes.

<table>
<thead>
<tr>
<th>Size of firm</th>
<th>Number of firms contacted</th>
<th>Number of firms surveyed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-1 1-2 2-5 5-10 ≥ 10</td>
<td>0-1 1-2 2-5 5-10 ≥ 10</td>
</tr>
<tr>
<td></td>
<td>yr. yrs. yrs. yrs. yrs.</td>
<td>yr. yrs. yrs. yrs. yrs.</td>
</tr>
<tr>
<td>0-3 employees</td>
<td>30 20 69 137 330</td>
<td>13 9 33 39 65 159</td>
</tr>
<tr>
<td>4-9 employees</td>
<td>7 14 56 80 103 260</td>
<td>2 8 29 47 61 147</td>
</tr>
<tr>
<td>10-19 employees</td>
<td>4 8 37 57 100 206</td>
<td>2 6 19 36 59 122</td>
</tr>
<tr>
<td>20-49 employees</td>
<td>2 3 28 55 91 179</td>
<td>1 3 20 25 59 108</td>
</tr>
<tr>
<td>≥50 employees</td>
<td>6 6 11 12 52 87</td>
<td>5 6 10 9 48 78</td>
</tr>
<tr>
<td>Missing data</td>
<td>27 7 21 16 28 99</td>
<td>5 2 6 6 8 27</td>
</tr>
<tr>
<td>Total</td>
<td>76 58 222 294 511 1161</td>
<td>28 34 117 162 300 641</td>
</tr>
</tbody>
</table>

The real survey often encountered difficulties. One of these concerned firms
which did not belong to the textile and clothing industry, and they amounted to
96. Specifically, 1 firm had relocated in another region, 21 firms had closed before
1991, and 74 were extraneous firms, i.e., they were carrying out economic activities
differing from the object of interest, though quite similar, such as the manufacture of
fabrics or related secondary activities, such as marketing and distribution. The firms
that had closed represent the death rate resulting from the delay in updating the
frame and the emigrated firms may be considered in the same manner as those that
had closed down, but handling the 74 misclassified firms was an uncertain matter.

a) They could be ignored and treated as nonrespondents, but a possible over­
estimation of the population parameters might be obtained.

b) It is plausible that while there are extraneous firms in the frame, perhaps
there are firms belonging to the target population which have been (mis)classified in
a different activity. Assuming that these two types of error have the same magnitude
and are evenly distributed over strata, the extraneous firms should be interviewed,
collecting only meaningful data, and included in the analyses. Although these assump­
tions seem reasonable, information about the phenomenon was affected by
great uncertainty, the size (number of firms) per strata was low, and the knowledge
 gained in the field prompted us to avoid this strategy.

c) An estimation of the actual number of firms in each stratum might be con­
sidered, but a possible underestimation of the population parameters could result.
However, given the specificity of the business activities excluded, the possibility of
underestimating the characteristics of interest was preferred. Therefore, the actual
number of firms in stratum $ij$ was calculated by

$$N_{A;ij1} = p_{ij1} N_{ij1} = \frac{n_{c;ij1}}{n_{c;ij1}} N_{ij1},$$

(3)

where the subscript 1 indicates the occasion number ($t = 1$), $N_{ij1}$ is the population
size obtained from the CERVED frame, $N_{A;ij1}$ is the estimated population size,
and $p_{ij1}$ is the sample proportion of firms belonging to the target population of
stratum $ij$ (determined by the ratio of the number of firms contacted and belonging
to stratum $ij$, $n_{c;ij1}$, to the total number of firms contacted in that stratum, $n_{c;ij1}$).
Introducing the correction for traceability and firms belonging to the population, the following was used

\[ w_{t;j;i} = \frac{1}{\pi_{t;j;i}} \frac{1}{\pi_{\text{trc};i;j}} \frac{1}{\pi_{\text{blg};i;j}} = \frac{N_{i;j}}{n_{i;j}} \frac{n_{i;j}}{n_{c;i;j}} \frac{n_{c;i;j}}{n_{E;i;j}} \]  

(4)

where \( \pi_{\text{trc};i;j} \) is the probability of being traced and \( \pi_{\text{blg};i;j} \) is the probability of belonging to the target population which were estimated by \( n_{c;i;j}/n_{i;j} \) and \( n_{E;i;j}/n_{c;i;j} \), respectively.

The probability of firms responding and participating in the survey, \( \pi_{r;i;j} \), could be estimated by the response rate which was calculated as the ratio of the number of firms interviewed or respondents, \( n_{r;i;j} \), to the number of firms selected, contacted and belonging to the target population, \( n_{E;i;j} \), in the weighting stratum \( i;j \)

\[ \pi_{r;i;j} = \frac{n_{r;i;j}}{n_{E;i;j}}. \]

The weight corrected to compensate also for firm nonresponse is given by

\[ w_{r;i;j} = \frac{1}{\pi_{t;i;j}} \frac{1}{\pi_{\text{trc};i;j}} \frac{1}{\pi_{\text{blg};i;j}} \frac{1}{\pi_{r;i;j}} = \frac{N_{i;j}}{n_{i;j}} \frac{n_{i;j}}{n_{c;i;j}} \frac{n_{c;i;j}}{n_{E;i;j}} \frac{n_{E;i;j}}{n_{r;i;j}}. \]  

(5)

This is also the simplest solution to compensate for nonresponses, although there are many other possible methods (Rubin, 1987). The estimators, however, become nonlinear and the respective variances may increase (Kish, 1990). Moreover, these corrections do not correlate with variability within the strata and they generally increase the variances of the estimates (Bethlehem and Keller, 1987; Potter, 1990).

In practice, it is useful to scale weights so that the mean value per firm which completed the interview is 1.0 because its application does not alter the size of the sample and does not expand the sample total to the population total (Verma, 1995).

There was also a necessity to add new units (referred to as replacement firms) to increase the size of the actual sample because in some strata the number of surveyed firms was disappointingly low. The inclusion probabilities of replacement firms was given by the sum of the probability of being drawn at the first stage, \( \pi_{i;j} \), and the probability of being drawn at the second stage of the same occasion, \( \pi_{i;j} + (1 - \pi_{i;j}) \pi_{rpl;i;j} \). Therefore, in each stratum \( i;j \), the weight for replacement firms (responding and participating in the survey) is given by

\[ w_{r;i;j} = \frac{1}{\pi_{i;j}} + \frac{1}{\pi_{\text{trc};i;j}} \frac{1}{\pi_{\text{blg};i;j}} \frac{1}{\pi_{r;i;j}} = \frac{N_{i;j}}{n_{i;j} + n_{rpl;i;j}} \frac{n_{i;j} + n_{rpl;i;j}}{n_{c;i;j} + n_{rpl;i;j}} \frac{n_{c;i;j} + n_{rpl;i;j}}{n_{E;i;j} + n_{rpl;i;j}} \frac{n_{E;i;j} + n_{rpl;i;j}}{n_{r;i;j}}. \]  

(6)

where \( n_{rpl;i;j} \) is the number of replacement firms selected in stratum \( i;j \) at time \( t = 1 \) in the case of selection carried out without replacement. Note that the evaluation of \( \pi_{\text{trc};i;j} \) also takes into account the fact that some replacement firms could not be traced.
3. Sampling procedure on subsequent occasions

If the purpose of a panel survey were solely that of a longitudinal analysis, it would be sufficient to simply follow the initial sample selected on the first occasion. However, as cross-sectional estimates were also of interest in the textile and clothing industry survey, it became necessary to update the sample at each occasion in order to represent new entrants to the population. The procedure adopted to obtain a sample that continues to be representative of the current population, is described below in the generic stratum $ij$.

1. At time $t = 1$, the initial sample with size equal to $n_{ij1}$ units was selected from the population of $N_{ij1}$ units.

2. At time $t = 2$, the elements selected on the previous occasion were included in the sample. More specifically, $m_{ij2}$ ($M_{ij2}$) denotes the number of units surviving at time $t = 2$ in the sample (population), i.e. existing on both occasions, and all the $m_{ij2}$ units were included in the sample.

3. Addition of a new sample to that of the surviving sampled units, $m_{ij2}$. Let $B_{ij2}$ be the number of non-existent units at time $t = 1$ and present at time $t = 2$ (newborn or immigrant firms) in the population, a sample of size $b_{ij2}$ was extracted only from $B_{ij2}$. The total sample was: $n_{ij2} = m_{ij2} + b_{ij2}$, while the total population was $N_{ij2} = M_{ij2} + B_{ij2}$. In determining the entity of $b_{ij2}$ the following rule was adopted:

$$b_{ij2} = \frac{n_{ij1}}{N_{ij1}}, \quad i.e., \quad b_{ij2} = \left[\frac{n_{ij1}}{N_{ij1}} \cdot B_{ij2} + 0.5\right], \quad (7)$$

where $[\cdot]$ indicates the integer part.

On subsequent occasions, the procedure starts from step 2 again and proceeds through step 3 until the data collection process over time is interrupted (Narain, Kathuria and Srivastava, 1987; Lalla, 1992). The generalisation for the generic phase $t$ uses the relationship for the dynamics of the population, $N_{ijt} = M_{ijt} + B_{ijt}$, where $M_{ijt} = N_{ijt-1} - D_{ijt}$ and $D_{ijt}$ denotes the number of units existing at time $t - 1$ and not found at time $t$ (dead or emigrated firms). The same relationships, with symbols in lowercase letters, hold for the sample. However, the rule used to determine $b_{ijt}$ could alter the precision of the estimation over time for new entrants to the population.

This scheme may be called a “replenished” panel survey because the original sample is “refreshed” at each new occasion. The units will remain in the sample for the entire duration of their life. The flows of the birth and death processes in the sample are independent, i.e., the firms added to the sample are selected only from the births and their number does not depend on the number of losses from deaths. Therefore, additions due to births and losses due to deaths in the sample are “attempts” to represent the population dynamics over time.

The newborn firms were directly obtained from the frame supplied by CERVED, but it was not always possible to check their origin. Thus, they were those units that had actually entered the market for the first time, those firms that had changed denomination and/or ownership, offshoots from already existing firms, and so on. Newborn firms also included immigrant firms, i.e., those firms with headquarters in another region and that had opened up a local unit for the first time. The firms sampled on the previous occasion, which were registered as births on the current occasion, were included in the sample to avoid the problem of “false” births.
Interview spacing was fixed at two years, but one year represents the minimum time interval between follow-up interviews required to detect budgetary items (such as turnover) and "general or structural" information (such as the type of final consumer, destination of the products, and so on).

3.1. Maintenance of the sampled units over time

As a rule, the firms belonging to the panel were not replaced if they had exited from the population because they had closed down. Indeed, it can be assumed that what happens in the sample approximately reflects what happens in the population. Firms that emigrated, i.e., those firms that had relocated or changed the type of economic activity, were considered in the same manner as those that had closed down. In general, transformations other than a relocation of headquarters or new type of economic activity (including new partners, separation from old partners, take-overs of an already existing unit, and so on) did not lead to that firm's exclusion from the sample. In particular, firms in the sample that merged with other firms that were not part of the sample, remained within the sample, "dragging" the resulting firm with them.

It was also assumed that the panel should retain every firm originating from a division (nonsample firms): one unit splits into two firms, a new firm is set up from the closure of one unit, one or more partners leave and the others remain and change the name of the firm, and so on. Finally, replacements for nonrespondents on subsequent occasions were carried out only if there was a risk of not having enough units in the respective strata, because they are presumably distributed unevenly over the strata and the sample would be biased in the long run.

This sampling strategy may, therefore, be considered a reasonable compromise between the constancy of the sample size and the representativeness of the sample of the population. In fact, if the ratio of sample size to population size (in each stratum) were maintained constant, the precision of the estimates would decrease steadily over time. Furthermore, the stratification by size of firms, limiting the variances of the number of employees in the strata determined by classes 1-4, should yield limited changes in the sample sizes in the strata resulting from re-application of the sampling procedure used on the first occasion. Therefore, this procedure should maintain almost the same allocation (with respect to the ratio \( n_{ij1}/N_{ij1} \)) as that of the first occasion (which is \( n_{ij1}/N_{ij1} \)).

3.2. The sample on the second occasion

The number of sampled firms surviving on the second occasion, \( m_{ij2} \), and the number of firms actually interviewed on the second occasion, \( n_{ij2} \), are reported in Table 2. Only 81 of the 91 large firms in the fifth size-class interval were interviewed because 2 firms had left, 2 firms were uncooperative (the same ones as on the previous occasion) and 6 firms did not belong to the target population, but 5 of these 6 firms were the same firms contacted on the previous occasion (see Section 2.2). Moreover, 3 firms appearing in the first and second columns in the left part of Table 2 had changed denominations and/or ownership, and CERVED had obviously registered them as newborn firms, while they were actually transformations. The first and second columns in the right section of Table 2 show the newborn firms, \( b_{ij2} \), included in the sample. The number of replacement firms per stratum involves only those strata defined by the last three age groups (third, fourth and fifth columns.
in the two sections of Table 2) and it may be obtained by subtracting the number of firms surviving in 1992 from the number of firms interviewed in 1992. Also included in Table 2 are the firms that exited from the market, losses by attrition, and transformations. Note, however, that the size of the present sample proved to be larger than that obtained on the first occasion because some units that had expressed their intentions not to participate, did co-operate after all, sending the completed questionnaire to the Regional Observatory or to the Small Entrepreneurs Association (by mail and long after the deadline).

The frame provided by CERVED gave rise to problems in this case as well. Thus, on the second occasion, 44 firms could not be traced, but 25 of the latter belonged to the initial sample established on the first occasion. Furthermore, 78 of the 96 units which did not belong to the target population on the first occasion were still present in the current frame. Moreover, it was ascertained that of the firms invited to take part in the survey for the first time, 22 did not belong to the target population: 7 among the newborn firms and 15 among the firms invited to participate in order to replace losses due to attrition. These difficulties caused by nonresponses and replacements, gave rise to different sampling fractions across strata.

Table 2 - Number of firms interviewed on the first occasion \((t_1 = 1990)\) and surviving on the second \((t_2 = 1992)\), \(n_{ijt_1}\), number of firms interviewed on the second occasion, \(n_{ijt_2}\), by size and age classes.

<table>
<thead>
<tr>
<th>Size of firm (1992)</th>
<th>(0^1-1) yr.</th>
<th>(1^1-2) yrs.</th>
<th>(2^1-5) yrs.</th>
<th>(5^1-10) yrs.</th>
<th>(\geq 10) yrs.</th>
<th>Tot.</th>
<th>(0^1-1) yr.</th>
<th>(1^1-2) yrs.</th>
<th>(2^1-5) yrs.</th>
<th>(5^1-10) yrs.</th>
<th>(\geq 10) yrs.</th>
<th>Tot.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-3 employees</td>
<td>2</td>
<td>22</td>
<td>44</td>
<td>69</td>
<td>137</td>
<td>14</td>
<td>22</td>
<td>26</td>
<td>54</td>
<td>75</td>
<td>191</td>
<td></td>
</tr>
<tr>
<td>4-9 employees</td>
<td></td>
<td>17</td>
<td>44</td>
<td>66</td>
<td>127</td>
<td>2</td>
<td>8</td>
<td>17</td>
<td>56</td>
<td>81</td>
<td>164</td>
<td></td>
</tr>
<tr>
<td>10-19 employees</td>
<td></td>
<td>9</td>
<td>36</td>
<td>57</td>
<td>102</td>
<td>1</td>
<td>10</td>
<td>43</td>
<td>69</td>
<td>123</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20-49 employees</td>
<td></td>
<td>8</td>
<td>19</td>
<td>67</td>
<td>94</td>
<td>1</td>
<td>8</td>
<td>21</td>
<td>70</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\geq 50) employees</td>
<td></td>
<td>1</td>
<td>12</td>
<td>11</td>
<td>45</td>
<td>1</td>
<td>2</td>
<td>14</td>
<td>11</td>
<td>53</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>Missing data</td>
<td></td>
<td>9</td>
<td>8</td>
<td>4</td>
<td>21</td>
<td>2</td>
<td>6</td>
<td>9</td>
<td>10</td>
<td>6</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>77</td>
<td>162</td>
<td>308</td>
<td>550</td>
<td>20</td>
<td>39</td>
<td>84</td>
<td>195</td>
<td>354</td>
<td>692</td>
<td></td>
</tr>
<tr>
<td>Dead firms</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Untraceable</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonrespondents</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>15</td>
<td>24</td>
<td>51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 3.3. The use of weights on subsequent occasions

Weighting in a longitudinal survey involves two types of dynamics: population dynamics and sample dynamics. The former refer to new entrants to and leavers from the population over time, while the latter refer to inclusions in and losses from the original sample over time. The sampling scheme tries to separate these dynamics through stratification by age. However, the addition of nonsample firms and the losses from nonresponse introduce several problems that require some explanation of their treatment in the estimates.
3.3.1. Panel nonrespondents

Nonrespondents can be grouped into those who fail to respond on the first occasion (initial nonrespondents or self-selection) and those who respond on the first occasion but fail to respond on one or more of the subsequent occasions of the panel survey for which they are eligible (panel nonrespondents). There are different causes that give rise to panel nonresponses: the firms belong to a category in which firms are more likely to drop out than others (Sobol, 1959); attrition, including refusal due to panel burden or disappearances that cannot be traced; retest reactivity, involving contamination of behaviour and attitudes of the selected interviews following the first one; reinterview laxity, due to indolence and fatigue of both the respondent and the interviewer; temporary nonresponse, due to either absences or refusals (Kish, 1989).

The weighting adjustments related to the panel nonrespondents generally modify the weights of panel respondents. One method consists of forming nonresponse adjustment cells and adjusting the weights by the inverses of the observed response rates in the cells. The cells are obtained through cross-classification of the responses from a set of variables measured on the first (previous) occasion, presuming that the latter are correlated with panel responses. The resulting sample size in each cell should be 30 or more; otherwise, it is necessary combine small cells (Chapman, Bailey, and Kasprzyk, 1986). The inverse of the observed response rate, \( \pi_{r;ijt} \), is the panel nonresponse adjustment for that cell:

\[
    w_{ijt} = \frac{1}{\pi_{ij1}} \cdot \frac{1}{\pi_{r;ijt}}
\]

where \( \pi_{ij1} \) includes the adjustment for nonresponse on the first occasion, but it does not include possible post-stratification adjustment. It is possible that \( \pi_{r;ijt} \) is correlated with \( \pi_{ij1} \) over cells, but they should be independent within a given cell \( ij \) and, therefore, the previous product is feasible.

Alternative methods for nonresponse adjustment can be classified into three groups (Rizzo, Kalton, and Brick, 1996). First, logistic regression: the weights determined on the first occasion are adjusted by the inverse of the probability that a unit responds on the subsequent occasion. The latter is obtained as a prediction from the (reduced main-effects) logistic regression model for each of the cells in the cross-classification of the predictor variables in that model. A mixed approach is also possible: in the cells containing 25 or more sampled firms, the adjustment is carried out through the inverse of the observed cell response rate. In cells containing less than 25 sampled firms, the adjustment is obtained by the inverse of the predicted response rate for the cell.

Second, the CHAID method: the panel nonresponse adjustment is the inverse of the observed response rate in the cells, which are defined as combinations of responses to the predictor variables that have the greatest discrimination with respect to the panel response rates, subject to the restriction that each cell should have at least 25 sampled firms (Kass, 1980). CHAID is an offshoot of AID (Automatic Interaction Detection) designed for a categorised dependent variable.

Third, the generalised raking method: the panel nonresponse adjustment is determined by "adapting" the marginal distributions of panel nonrespondents for each predictor variable (obtained using the adjusted weights) to the corresponding distributions for respondents and nonrespondents combined (obtained using the original
weights of the first occasion). Therefore, the original weights are modified in order to satisfy marginal constraints, minimising the distance between the original and adjusted weights. CALMAR software may be used to carry out the raking adjustment (Deville, Särndal, and Sautory, 1993).

As the sample was not large, the choice of auxiliary variables could be subject to some limitations and will represent a still more important stage than the choice of the weighting methods. In fact, when many auxiliary variables are included in the model used to adjust the weights, the results may fluctuate substantially (Rizzo, Kalton, and Brick, 1996).

3.3.2. The replacement firms

The replacement firms selected to counterbalance nonresponses and other losses should contribute only to increasing the precision of the cross-sectional estimates for the occasions on which the replacements take place. They might be considered as an "independent" sample drawn from the population of surviving firms which had not been included in the sample, $M_{ijt} - m_{cijt}$ (where $m_{cijt}$ is the number of firms contacted at time $t = 1$ and surviving at time $t$). In the stratum $ij$, their inclusion probabilities would be given by the sum of the probability of being drawn on the first occasion, $\pi_{ij1}$, and the corresponding probability of being drawn on the second occasion, $\pi_{ij1} + (1 - \pi_{ij1}) \pi_{rpl;ij2}$. If replacement firms were also selected on the first occasion in stratum $ij$, the probability of replacement firms being drawn on the second occasion would be

$$[\pi_{ij1} + (1 - \pi_{ij1}) \pi_{rpl;ij1}] (1 - \pi_{rpl;ij2}) + \pi_{rpl;ij2}.$$

This probability is easy to compute for populations which do not migrate from one stratum to another. Otherwise, the calculation becomes tedious because the selection probability of firms belonging to stratum $kl$ on the previous occasion, and $ij$ on the current occasion, should be given by the previous equation in which $\pi_{rpl;ij1}$ is replaced by $\pi_{rpl;kl1}$. However, if nonresponses and other losses are not random, as seems to be the case in the present survey, serious biases could be introduced into the sampling estimates of the target parameters in the long run.

3.3.3. The nonsample firms: merging and splitting

Mergers and splits give rise to nonsample firms which might be excluded from the weighted analysis because it is very difficult to determine their selection probabilities precisely. In fact, let us consider a firm consisting of a sample firm which merges with a nonsample firm on the current occasion. The probability of the resulting firm being included in the current sample is the sum of the probability that the sample firm was selected on the first occasion ($t = 1$) and the corresponding probability of the nonsample firm being included. The former is known, whereas the latter is unknown. A rough approximation for the latter could be obtained from the probability of the merging occurring. Another possible strategy, similar to the previous one, is the following: i) the "new" firms formed by mergers with old firms (which existed at the previous survey time) are excluded from the newborn population; ii) for all sample firms involved in a merger, the weights are proportional to the number of old firms involved in the merger. Difficulties arise in ascertaining the "new" firms set up through mergers, because there is not enough information in the CERVED frame.
Mergers require a more in-depth investigation of their effect on the estimates, even if they would presumably represent a negligible fraction of earlier survey results. In fact, it is true that if two firms existing on the first occasion merge before the second occasion, the merged firm has two chances of selection on the second occasion. However, adjustment for this in the analysis may not be performed straightforwardly because in practice, merging is not generally equal to the sum of the characteristics of the mergers, but involves their reduction as well as their increase. When merging concerns sampled firms only and given the specific design, at a fixed age-class interval, the following could represent some possible rules.

1) If two or more firms belonged to the same stratum \( ij \) at time \( t-1 \), merged between \( t-1 \) and \( t \), and the resulting merged firm still belongs to the stratum \( ij \), then the selection probability for the latter would be equal to that of one of them.

Example 1. Let us consider only one stratum, and drop the \( ij \) subscripts for simplicity. Suppose at time \( t = 1 \), \( N_1 = 1000 \) and \( n_1 = 100 \). The selection probability is \( \pi_{1k} = 0.1 \) for generic firm \( k \). Suppose at time \( t = 2 \), \( m_2 = 95 \), i.e., "90 original firms" plus "5 merged firms". Each merged firm is formed by two sampled firms existing at \( t = 1 \). In this case, it is not necessary to change the selection probability of the merged firms because, assuming that the sample is representative of the population, it will be \( M_2 = \sum_{k=1}^{m_2} 1/\pi_{1k} = 950 \), i.e., 50 merged firms should be present in the population.

2) If two or more firms belonging to different strata at time \( t-1 \), merged between \( t-1 \) and \( t \), and whatever the resulting stratum may be for the merged firm, then the selection probability for the merged firm would be equal to that of the merger belonging to stratum \( ij \), in which \( N_{ij} \) provides an upper boundary for the total number of merged firms. However, in one stratum (or more than one), the estimation of the population parameters could be biased.

Example 2. Let us consider only three strata, and drop the \( j \) subscript for simplicity. The population sizes, \( N_i \), the selection probabilities, \( \pi_{il} \), and the sampling sizes, \( n_i \), at time \( t = 1 \), are reported in Table 3, for \( i = A, B, C \). Three different cases are considered in Table 3, but only the first is discussed here because the others are interpreted in the same way. Each merged firm belongs to stratum \( C \) and is formed by three sampled firms existing at \( t = 1 \): one belonging to stratum \( B \) and two belonging to stratum \( A \). The total number of merged firms in the sample is 20. Assuming that the sample is representative of the population, the maximum number of merged firms can be 100 only, i.e., \( n_{B1}/\pi_{B1} \). In fact, \( n_{A1}/\pi_{A1} = 400 \) implies 200 merged firms (on the basis of the adopted rule), but this is impossible. Therefore, in stratum \( A \), the actual population is \( N_{A2}^{(1)} = 1800 \), but the estimation on the basis of \( n_{A2}^{(1)} = 1600 \), i.e., the selection probability for firms belonging to stratum \( A \) fails to estimate the population size.

Considering the size-class and age-class intervals simultaneously, the rules could remain the same, assuming that the merged firm belongs to the same age-class as the oldest firm.

Splitting gives rise to genuine nonsample firms. The difficulties in handling them may be partially resolved when the splits are retained in the sample, as is the case in this study, because they could reflect the splits that occurred in the population. However, as splitting concerns only one statistical unit, a simple solution for weights could be that of assigning to the split firms, the same selection probability as that of the original (disappearing) firm. Three different cases of splitting firms are reported.
in Table 4: (1) all the split firms belong to a stratum which is the same as that of the original (disappearing) firm; (2) some of the split firms belong to a stratum which is the same as that of the original (disappearing) firm and some others belong to a stratum which is different from that of the latter; (3) all the split firms belong to a stratum which is different from that of the original (disappearing) firm. The values of \( N_i^{(l)} \) (for \( l = 1, 2, 3 \)) are obtained by weighting the split firms by the weight of the original firm. Again, the assumption is that what happens in the sample, also happens in the population with the same proportion.

Table 3 – Some cases of merging occurring between the first and second occasions.

<table>
<thead>
<tr>
<th>Strata</th>
<th>( N_i^{(1)} )</th>
<th>( n_i^{(1)} )</th>
<th>( mrg^{(1)} )</th>
<th>( n_i^{(2)} )</th>
<th>( N_i^{(2)} )</th>
<th>( mrg^{(2)} )</th>
<th>( n_i^{(3)} )</th>
<th>( N_i^{(3)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2000</td>
<td>0.1</td>
<td>200</td>
<td>-40</td>
<td>160*</td>
<td>180*</td>
<td>-20</td>
<td>180*</td>
</tr>
<tr>
<td>B</td>
<td>1000</td>
<td>0.2</td>
<td>200</td>
<td>-20</td>
<td>180</td>
<td>900</td>
<td>-40</td>
<td>160</td>
</tr>
<tr>
<td>C</td>
<td>500</td>
<td>0.4</td>
<td>200</td>
<td>+20</td>
<td>220</td>
<td>600</td>
<td>+20</td>
<td>220</td>
</tr>
</tbody>
</table>

\( n_i^{(1)} \) and \( n_i^{(2)} \) denote the number of merged firms in stratum \( i \) at the first and second occasions, respectively. \( N_i^{(1)} \) and \( N_i^{(2)} \) are the estimated number of population firms in stratum \( i \) at the first and second occasions, respectively. \( mrg^{(1)} \) and \( mrg^{(2)} \) are the selection probability of merging in stratum \( i \) at the first and second occasions, respectively.

\( (1) \) MF: Merged Firm; \( F \): Firm \( \pi_{mrg:C2} = 0.2 \)

\( (2) \) MF: Merged Firm; \( F \): Firm \( \pi_{mrg:C2} = 0.2 \)

\( (3) \) MF: Merged Firm; \( F \): Firm \( \pi_{mrg:A2} = 0.4 \)

* The selection probability fails to estimate population size.

Table 4 – Some cases of splits occurring between the first and second occasions.

<table>
<thead>
<tr>
<th>Strata</th>
<th>( N_i^{(1)} )</th>
<th>( n_i^{(1)} )</th>
<th>( split^{(1)} )</th>
<th>( n_i^{(2)} )</th>
<th>( N_i^{(2)} )</th>
<th>( split^{(2)} )</th>
<th>( n_i^{(3)} )</th>
<th>( N_i^{(3)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2000</td>
<td>0.1</td>
<td>200</td>
<td>200</td>
<td>2000</td>
<td>+20</td>
<td>220</td>
<td>2100</td>
</tr>
<tr>
<td>B</td>
<td>1000</td>
<td>0.2</td>
<td>200</td>
<td>(20+20)→-B 220</td>
<td>1100</td>
<td>20→-B</td>
<td>200</td>
<td>1000</td>
</tr>
</tbody>
</table>

3.3.4. Population and sample dynamics

Population dynamics between strata (growth in firms) seem not to affect the weights if the stratification variable does not notably modify its distribution over strata. In fact, once a firm is in the sample, its sample weight is determined and, whatever its current stratum may be, it will be held constant over time. An example of transitions of firms from one stratum to another over time (from \( t = 1 \) to \( t = 2 \)) are reported in Table 5 which shows, for instance, that stratum \( A \) contains 200 firms at time \( t = 1 \), \( n_{A1} \), sampled from a population \( N_{A1} = 2000 \) (the same selection probabilities as in Table 3). At time \( t = 2 \), stratum \( A \) contains 220 sampled firms, of which 160 firms belonged to \( A \) (\( \pi_{A1} = 0.1 \)), 40 firms belonged to \( B \) (\( \pi_{B1} = 0.2 \)), and 20 firms belonged to \( C \) (\( \pi_{C1} = 0.4 \)) at time \( t = 1 \). The weights for the first occasion yielded unbiased estimates for the number of population firms, \( N_{i,t} \) for \( i = A, B, C \). However, the estimates could be biased when rare events or rare transitions occur in the population or in the sample because it is difficult to obtain precise estimates of their magnitude. The population of firms common to both occasions (the current and the previous one) could also be over-represented or under-represented when the
number of deaths in the population, $D_{ijt}$, or the transitions between strata are not accurately represented by the sample. These problems may be corrected by applying the method of post-stratification (Cochran, 1977; Holt and Smith, 1979). A rough correction of the weights might be given by the ratio of the current population in each stratum to the sample firms in the corresponding strata, which is similar to post-stratification.

The use of optimal allocation gives rise to another difficulty in maintaining the panel sample representative of the population over time. The $m_{ij2}$ firms of the matched portion in the sample on the second occasion have a selection probability of $\pi_{ij1k}$ for each firm $k$ in stratum $ij$. However, some firms could have changed their characteristics over time, such as number of employees, turnover, type of product, destination of goods, and so on. They would now belong to different strata and the selection probabilities, $\pi_{ij1k}$, of these firms will not be representative of their actual probability of being included in the sample at the present time $t = 2$. In other words, if a new sample of firms is selected at $t = 2$, the selection probabilities $\pi_{ij2k}$ of the firms participating on both occasions would differ from the $\pi_{ij1k}$ attributed to them at $t = 1$. These difficulties may be overcome by applying the method of post-stratification once again.

![Table 5 - Transitions between strata from one occasion to another.](image)

### 4. Methods of estimation

On the second occasion, as well as on the subsequent occasions, the available measured characteristics refer to three subsamples: a) newborn firms, $b_{ijt}$, which were selected on the current occasion from the population of newborn firms in each stratum $ij$ for $j \leq 2$; b) replacement firms, $n_{rpl,ijt}$, which were selected on the current occasion from the population of surviving firms in each stratum $ij$ for $j > 2$; c) surviving firms (in the sample), $m_{ijt}$, which were selected on the previous occasion and were still surviving at the current time in each stratum $ij$ for $j > 2$. The objectives of the data analysis were the estimation of population parameters, net and gross changes at a given point in time. Procedures for estimating the characteristics of a population in a panel or rotating panel survey, were first proposed by Jessen (1942), Yates (1949), Patterson (1950), and subsequently discussed by many others, among whom Eckler (1955), Hansen et al. (1955), Kulldorff (1963), Rao and Graham (1964), Gurney and Daly (1965), Sen (1973), Graham (1973), and Bellhouse (1991). In the present context, however, it was necessary to include another source of variability in the expressions regarding the variance of the estimators, because of the
need to estimate the size of the population owing to errors in the CERVED frame. Considering the correction for finite populations, the variance of the estimator of the target population in stratum \( ij \), at a given time \( t \), was given by

\[
V[N_{A;ijt}] = V[N_{ijt} p_{ijt}] = N_{ijt}^2 \frac{p_{ijt} q_{ijt}}{n_{c;ijt} - 1} \frac{N_{ijt} - n_{c;ijt}}{N_{ijt}},
\]

where \( q_{ijt} = 1 - p_{ijt} \).

The estimation of \( N_{A;ijt} \) was problematic because there were firms in the frame with an incorrect industry code and thus not belonging to the target population, as well as firms belonging to the target population which were not included in the frame. As noted in Section 2.3, exclusions generally regarded firms carrying out activities related to the textile and clothing industry, and a possible "weak" underestimation arising from this correction was preferred to an overestimation. Therefore, \( N_{A;ijt} \) was estimated through the sample proportion, \( p_{ijt} \), and was used in estimating the population parameters.

4.1. Cross-sectional estimates

The estimator of the population total for the characteristic of interest \( Y \), e.g. the number of employees or turnover, results from the sum of the estimators by stratum, \( \hat{Y}_t = \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{Y}_{ijt} \), where the general expression for the addenda is

\[
\hat{Y}_{ijt} = \sum_{k} w_{ijk} y_{ijt},
\]

in which the sum is over the \( k \) population values in stratum \( ij \) and \( w_{ijk} \) is a random variable that takes value \( w_{ijk} = 0 \), if the \( k \)-th firm is not in the sample. The expected value of \( \hat{Y}_{ijt} \) is

\[
E(\hat{Y}_{ijt}) = \sum_{k} E(w_{ijk}) y_{ijt},
\]

and \( \hat{Y}_{ijt} \) is an unbiased estimator of \( Y_{ijt} \) in any weighting scheme having \( E(w_{ijk}) = 1 \) for all \( k \).

For sample data, the estimator becomes

\[
\hat{Y}_{ijt} = \sum_{k} w_{ijk} y_{ijt},
\]

where the sum is over the \( k \) sample observations in stratum \( ij \). However, in the present case, the weights \( w_{ijk} \) include a function involving the product \( N_{ijt} p_{ijt} \), where \( p_{ijt} \) is the sample proportion of firms belonging to the target population and it is also a random variable implying nonlinearity in the estimators. Therefore, the usual weighting methods for cross-sectional estimates from longitudinal data (Lavallée, 1995; Kalton and Brick, 1995) were not applicable and the evaluation of the variance of \( \hat{Y}_t \) was achieved through two different strategies to compare the gain in precision that resulted. The sample variance of the sample mean per stratum, \( s_{2;ijt}^2 \), and \( p_{ijt} \), were used in both strategies.

First case (A). The previous survey and the three subsamples were ignored, and for every \( k \) sample observations in stratum \( ij \) at time \( t \), the weight was calculated by

\[
w_{ijt} = N_{ijt} p_{ijt} / n_{r;ijt},
\]

where \( n_{r;ijt} \) is the number of firms participating in the survey (respondents). Thus, the resulting weights were roughly adjusted to compensate for nonresponses, noncoverage, and changes that occurred in the population over time. Therefore, the estimate of the population total in each stratum was given by:

\[
\hat{Y}_{ijt} = \frac{N_{ijt} p_{ijt}}{n_{r;ijt}} y_{ijt} = \frac{N_{A;ijt}}{n_{r;ijt}} y_{ijt}.
\]

Considering the finite population correction and bearing in mind that the weight contained a random variable which represented the estimate of the total units per stratum, it was possible to extend summation to \( I \) and it thus follows that the
variance of $\hat{Y}_t$ is

$$V[\hat{Y}_t] = \sum_{i=1}^{I} \sum_{j=1}^{J} V \left[ \frac{N_{ijt} \ p_{ijt}}{n_{r;i.t}} \ y_{ijt} \right]$$
 $$= \sum_{i=1}^{I} \sum_{j=1}^{J} \left[ N_{ijt}^2 \ \bar{y}_{ijt}^2 \ \frac{p_{ijt} q_{ijt}}{n_{c;ijt} - 1} \ \frac{N_{ijt} - n_{c;ijt}}{N_{ijt}} \right]$$
 $$+ \ N_{A;ijt}^2 \ \frac{s_{ijt}^2}{n_{r;i.t}} \ \frac{N_{A;ijt} - n_{r;i.jt}}{N_{A;ijt}}.$$  \quad (9)

Second case (B). The previous survey and the three subsamples were taken into account, the replacement firms were considered as an independent sample drawn from the population of surviving firms, and in each stratum $ij$, the weight was calculated as above. To obtain the expression of $Y_t$, the following terms were defined: $\bar{y}_{bijt}$ is the mean of $Y$ in stratum $ij$ for the sampled units $bijt$, selected from the elements entering at time $t$, $B_{ijt}$, and $\bar{y}_{m;ijt}$ is the mean of $Y$ in the sample of matched units, $m_{ijt}$, present on both occasions $(t, t-1)$, and $\bar{y}_{r;ijt}$ is the mean of $Y$ in the sample of replacements units, $n_{r;i.jt}$. Considering that the newborn firms belong to the strata defined by $j \leq 2$, the estimate of $Y_{ijt}$ for $j > 2$ can be roughly approximated by weighting two “independent” estimates: $\bar{y}_{m;ijt}$ and $\bar{y}_{r;ijt}$. Therefore, the estimate of the population total in each stratum was given by

$$\hat{Y}_{ijt} = B_{A;ijt} \ \bar{y}_{bijt} + \left[ (1 - \psi_{ijt}) \ M_{A;ijt} \ \bar{y}_{m;ijt} + \psi_{ijt} \ M_{A;ijt} \ \bar{y}_{r;ijt} \right],$$  \quad (10)

where $M_{A;ijt} = M_{ijt} \ p_{ijt}$ and $\psi_{ijt}$ is the weight that was set as equal to the proportion of replacement firms in stratum $ij$: $\psi_{ijt} = n_{r;p;ijt}/n_{ijt}$. In the matched portion of the sample, $\bar{y}_{m;ijt}'$ was obtained through a double-sampling regression estimate, in which the auxiliary variate $x_{ijtk}$ (of the $k$-th firm) was the value of $y_{ijt-1,k}$, implying $\bar{y}_{m;ijt}' = \bar{y}_{m;ijt} + \beta(\bar{y}_{ijt-1} - \bar{y}_{m;ijt-1})$, where $\beta$ is the regression coefficient (see Cochran, 1977, pp. 346–349). This estimator is similar to a composite estimator, which combines the matched sample estimator with the unmatched sample estimator (Fuller, 1990; Wolter, 1979), and the variance of $\hat{Y}_t$, the estimator of the population total, has the following complex and extended expression, including the finite population correction

$$V[\hat{Y}_t] = \sum_{i=1}^{I} \sum_{j=1}^{J} \left[ B_{ijt}^2 \ \bar{y}_{bijt}^2 \ \frac{p_{ijt} q_{ijt}}{n_{c;ijt} - 1} \ \frac{M_{ijt} - n_{c;ijt}}{M_{ijt}} \right]$$
 $$+ \ \sum_{i=1}^{I} \sum_{j=3}^{J} \left[ (1 - \psi_{ijt})^2 \ M_{ijt}^2 \ \bar{y}_{m;ijt}^2 \ \frac{p_{ijt} q_{ijt}}{n_{c;ijt} - 1} \ \frac{M_{ijt} - n_{c;ijt}}{M_{ijt}} \right]$$
 $$+ \ M_{A;ijt}^2 \ \left[ \frac{s_{ijt}^2}{M_{ijt}} (1 - \rho_{ijt}^2) \ \frac{M_{A;ijt} - n_{ijt}}{M_{A;ijt}} + \rho_{ijt}^2 \ V[\bar{y}_{ij,t-1}] \right]$$
 $$+ \ \psi_{ijt}^2 \ M_{ijt}^2 \ \bar{y}_{r;p;ijt}^2 \ \frac{p_{ijt} q_{ijt}}{n_{c;ijt} - 1} \ \frac{M_{ijt} - n_{c;ijt}}{M_{ijt}}$$
 $$+ \ M_{A;ijt}^2 \ \frac{s_{r;p;ijt}^2}{n_{r;p;ijt}} \ \frac{M_{A;ijt} - n_{r;p;ijt}}{M_{A;ijt}}.$$  \quad (11)
In order to determine the variance of the total, at a given time \( t \), the variance \( V[\hat{Y}_{ij,t-1}] \) must be available. Therefore, \( V[\hat{Y}_t] \) is further complicated for \( t > 2 \). However, this procedure is only one possible approximation of the estimate of the variance which otherwise requires a more complex procedure (Cicchitelli, Herzel and Montanari, 1992, pp. 439-467; Särndal, Swensson and Wretman, 1992).

As an example, Table 6 reports the estimates of the number of employees and the turnover resulting from the two cases (A) and (B). The former yielded values lower than the latter. The respective standard errors were calculated with expressions (9) and (11). The latter considers the information derived from the panel, i.e., from the proportion of firms that took part in both surveys and it generally yielded an increase in the precision of the estimates by 3-30%, with the respect to the former. In fact, the variance of the estimate due to classification errors in the frame varied from one stratum to another, but it was often of the same order of magnitude as the other terms arising from the random sampling variation in equations (9) and (11). However, in some strata, the variance calculated by equation (11) was greater than that resulting from equation (9) because the sample mean and the sample variances for the current occasion differed markedly from the corresponding values observed on the previous occasion. In fact, if the changes in the stratification (and other) characteristics occurring in a stratum are marked, for example, one or more firms greatly increase the number of employees and turnover occurring between the two occasions, the uncertainty (and hence the variance) should increase, as it did.

In the matched portion of the sample, \( \hat{Y}_{mi,ij,t} \) was also estimated through an average of \( \hat{Y}_{m,ij,t} \) and \( \hat{Y}_{mi,ij,t} \), where the latter was obtained by an "auto-regressive model" implying: \( \hat{Y}_{mi,ij,t} = \phi_{ij} \hat{Y}_{mi,ij,t-1} \), where \( \phi_{ij} \) is the regression coefficient between the variates \( y_{mi,ij} \) and \( y_{mi,ij,t-1} \). In calculating the variance of the estimator of the population total, if \( \phi_{ij} \) is obtained from the frame, it should be held constant; if not, it should be treated as a variate. The expression of the variance was omitted because it is too extended, although it is similar to the previous equation (11). The corresponding estimates of the number of employees and turnover, as well as their standard errors, which were generally slightly lower than those reported in Table 9, have been left out for the sake of brevity. Finally, it should be noted that the estimation of the population parameters were obtained using some approximated procedures because the owners of the survey data have not yet permitted access to the files.
Table 6 – Estimates of the total number of employees and the turnover of the population, $\hat{Y}_{ij2}$, by size and age classes, obtained ignoring the sampling design and from equation (10), with respective standard errors (S.E.) calculated from equations (9) and (11).

<table>
<thead>
<tr>
<th>Size of firm</th>
<th>Estimates of the total no. of employees</th>
<th>Estimates of turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-1</td>
<td>1-2</td>
</tr>
<tr>
<td></td>
<td>yr.</td>
<td>yrs.</td>
</tr>
<tr>
<td>0-3 employees</td>
<td>364</td>
<td>848</td>
</tr>
<tr>
<td>$\hat{Y}_{ij2}$ from (10)</td>
<td>364</td>
<td>848</td>
</tr>
<tr>
<td>S.E. from (9)</td>
<td>57</td>
<td>161</td>
</tr>
<tr>
<td>S.E. from (11)</td>
<td>57</td>
<td>161</td>
</tr>
<tr>
<td>4-9 employees</td>
<td>86</td>
<td>223</td>
</tr>
<tr>
<td>$\hat{Y}_{ij2}$ from (10)</td>
<td>86</td>
<td>223</td>
</tr>
<tr>
<td>S.E. from (9)</td>
<td>9</td>
<td>31</td>
</tr>
<tr>
<td>S.E. from (11)</td>
<td>9</td>
<td>31</td>
</tr>
<tr>
<td>10-19 employees</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td>$\hat{Y}_{ij2}$ from (10)</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td>S.E. from (9)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S.E. from (11)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20-49 employees</td>
<td>33</td>
<td>678</td>
</tr>
<tr>
<td>$\hat{Y}_{ij2}$ from (10)</td>
<td>33</td>
<td>680</td>
</tr>
<tr>
<td>S.E. from (9)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S.E. from (11)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>≥50 employees</td>
<td>87</td>
<td>1861</td>
</tr>
<tr>
<td>$\hat{Y}_{ij2}$ from (10)</td>
<td>87</td>
<td>1861</td>
</tr>
<tr>
<td>S.E. from (9)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S.E. from (11)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Missing data</td>
<td>80</td>
<td>314</td>
</tr>
<tr>
<td>$\hat{Y}_{ij2}$ from (10)</td>
<td>80</td>
<td>314</td>
</tr>
<tr>
<td>S.E. from (9)</td>
<td>30</td>
<td>203</td>
</tr>
<tr>
<td>S.E. from (11)</td>
<td>30</td>
<td>203</td>
</tr>
<tr>
<td>Total</td>
<td>650</td>
<td>3426</td>
</tr>
<tr>
<td>$\hat{Y}_{ij2}$ from (10)</td>
<td>650</td>
<td>3426</td>
</tr>
<tr>
<td>S.E. from (9)</td>
<td>65</td>
<td>261</td>
</tr>
<tr>
<td>S.E. from (11)</td>
<td>65</td>
<td>262</td>
</tr>
</tbody>
</table>

4.2. Longitudinal estimates

In principle, the methods of the generalised linear multivariate model (Hsiao, 1986) can be applied to longitudinal data sets, even if an adaptation of the regression model to their specific structure is necessary:

$$y_i = \alpha_i + X_i \beta + \epsilon_i,$$

where, given $n$ panel firms and $T$ number of occasions, $y_i$ is a $(T \times 1)$ vector, $X_i$ is a $(T \times K)$ matrix of observation on $K$ explanatory variables for the $i$-th firms, $\beta$ is a $(K \times 1)$ vector of parameters to be estimated, $\epsilon_i$ is a $(K \times 1)$ vector of disturbances such that $E[\epsilon_i] = 0$. The individual effect is represented by $\alpha_i$, which
is taken to be constant over time, and specific to the individual cross-sectional firm, \( i \). Assuming \( \alpha_i \)'s to be the same across units, the ordinary least squares method provides consistent and efficient estimates of \( \alpha \) and \( \beta \). On the contrary, a simple formulation of the model assumes that differences across firms are represented by differences in \( \alpha_i \)'s (fixed effect). Therefore, the latter are parameters to be estimated in the model which is referred to as the least squares dummy variable (Greene, 1991):

\[
\begin{pmatrix}
 y_1 \\
 y_2 \\
 \vdots \\
 y_n \\
\end{pmatrix} = \begin{pmatrix}
 i & 0 & \ldots & 0 \\
 0 & i & \ldots & 0 \\
 \vdots & \vdots & \ddots & \vdots \\
 0 & 0 & \ldots & i \\
\end{pmatrix} \begin{pmatrix}
 \alpha_1 \\
 \alpha_2 \\
 \vdots \\
 \alpha_n \\
\end{pmatrix} + \begin{pmatrix}
 X_1 \\
 X_2 \\
 \vdots \\
 X_n \\
\end{pmatrix} \beta + \begin{pmatrix}
 \epsilon_1 \\
 \epsilon_2 \\
 \vdots \\
 \epsilon_n \\
\end{pmatrix} \\
\tag{13}
\]

where \( y_i \) and \( X_i \) are the \( T \) observations for the \( i \)-th firm, \( i \) is the \((T \times 1)\) vector of 1s, and \( \epsilon_i \) is the corresponding \((T \times 1)\) vector of disturbances.

With longitudinal data sets, it might be more appropriate to view the individual specific constant term as randomly distributed across cross-sectional firms (random effect). The reformulation of the model is

\[
y_i = \beta_0 + X_i \beta + \nu_i + \epsilon_i, \\
\tag{14}
\]

where \( \beta_0 \) is the constant term and \( \nu_i \) is the random disturbance representing the collection factors of the \( i \)-th observation, not in the regression, that are specific to the firm and are constant over time. For these \( T \) observations, let \( w_{it} = \nu_i + \epsilon_{it} \), then

\[
\begin{align*}
E[w_{it}] &= 0 \\
E[w_{it}^2] &= \sigma_\nu^2 + \sigma_\epsilon^2 \\
E[w_{it}w_{is}] &= \sigma_\nu^2 \text{I} \\
E[w_{it}w_{it}'] &= \Omega = \sigma_\nu^2 \text{I}' + \sigma_\epsilon^2 \text{I}.
\end{align*}
\tag{15}
\]

Assuming that the observations are independent, the disturbance covariance matrix for the full \( nT \) observations is

\[
V[w] = \begin{bmatrix}
\Omega & 0 & \ldots & 0 \\
0 & \Omega & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \Omega
\end{bmatrix},
\]

the structure of which is very simple.

4.2.1. Modelling panel nonresponses

Panel nonresponse involves not only a low statistical power, due to small sample size, but also the risk that nonresponse is not random (or selective) and the conclusions based on the analyses conducted on such a biased sample cannot easily be generalised for the population.

To account for a nonresponse pattern, on the one hand, it is possible to use a model based on a discrete-time Markov chain (Taris, 1996). On the other hand, it is possible to estimate the effect of nonresponse on estimates of the relationship between a dependent variable \( y_{it} \) (such as turnover) and a set of explanatory variables
such as the number of employees, type of firm, product, customer, destination of the products, final consumer), for $i = 1, \ldots, n$, the equation will be

$$y_i = \beta_0 + X_i \beta + \bar{X}_i \theta + \alpha_i + \epsilon_i.$$ (16)

As in the previous equation, $\alpha_i$ represents the unobserved firm characteristics and it is uncorrelated with $X_i$, while $\bar{X}_i$ denotes the time average of $X_{it}$. The latter is introduced in equation (14) to correct for possible correlations between $\nu_i$ and $X_i$ (Nijman and Verbeek, 1992).

To account for the effect of panel nonresponses, the dummy variable, $r_{it}$, is defined to indicate whether firm $i$ is asked to cooperate in period $t$ and has participated in the survey (0 if not, 1 otherwise). It is assumed that the inclusion of a firm in the sample is independent of the disturbances $\alpha_i$ and $\epsilon_i$ in the previous equation, but dependence on the exogenous variables is not excluded. Moreover, given that a firm was selected in the sample, it is possible to represent its decision to cooperate or not in the survey through a response equation. The dependent variable, $y_{it}$, is observed if a latent variable $r_{it}^*$ is non-negative. The latter could be explained by a (latent) regression equation:

$$r_{it} = \begin{cases} 0 & \text{if firm } i \text{ did not participate} \\ I(r_{it}^* > 0) & \text{if firm } i \text{ did participate} \end{cases}$$

where $\xi_i$ is an individual specific effect independent of $X_i$, and $z_i$ contains factors influencing nonresponse, but not influencing the dependent variable $y_i$. For example, it contains the dummy variable $r_{i,t-1}$ indicating whether a firm participated in the previous period or not (Nijman and Verbeek, 1992).

The models for $y_i$ and $r_i$ constitute a system of regression equations characterising sample selection bias (Heckman, 1979) in which the disturbances, $\epsilon_i$ and $\eta_i$, are normally distributed according to

\[
\begin{pmatrix}
\mathbf{i} \alpha_i + \epsilon_i \\
\mathbf{i} \xi_i + \eta_i
\end{pmatrix} \sim \mathcal{N}
\left[
\begin{pmatrix}
0 \\
0
\end{pmatrix};
\begin{pmatrix}
\sigma^2_{\epsilon} \mathbf{I} + \sigma^2_{\epsilon} \mathbf{I} & \sigma^2_{\epsilon \eta} \mathbf{I}
\\
\sigma^2_{\epsilon \eta} \mathbf{I} & \sigma^2_{\eta} \mathbf{I} + \sigma^2_{\xi} \mathbf{I}
\end{pmatrix}
\right],
\]

and this vector of disturbances is independent of $x_{it}$ ($\forall i, t$). In this framework, nonresponse is random and no selection bias occurs in $y_i$ if the unobserved determinants of response are uncorrelated with the unobserved determinants of $y_i$, i.e., if $\sigma_{\epsilon \eta} = \sigma_{\epsilon \xi}$.

5. Conclusions

After carrying out the survey on three occasions, the strategy described above seems satisfactory for the objectives of the survey, which were to provide information on the textile and clothing industry, an estimation of the level of the population characteristics, and net and gross change. The stratification according to age, which represents another way to introduce the fourth dimension, made it easier to separate the newly established firms from the surviving firms. Although the frame supplied by CERVED showed some imperfections, it is the only one that included all firms.
The results obtained appear to be reliable and the sampling scheme adopted may be easily applied to other regions. Obviously, the construction of a high-quality frame is desirable, but high costs and much time would be involved. However, elimination of the imperfections of the frame may be almost illusory because changes in population can occur at any time. There is an unavoidable delay in collecting and updating information, and only data on a few characteristics can be collected accurately by administrative offices.

This is a real survey conducted under typical practical constraints and it is thus affected by the usual difficulties: nonresponse, replacements, and noncoverage. Most of problems encountered in this study have been discussed and some possible corresponding solutions have been applied or proposed to solve them in an attempt to identify the strategy most appropriate to and consistent with the present situation.

Our aim was to obtain a maximum amount of information with margins of error and costs remaining within acceptable limits.

6. References


24. Fernando Vianello [1987] “Effective Demand and the Rate of Profit. Some Thoughts on Marx, Kaleski and Sraffa”, pp. 41
43. Giovanna Procacci [1989] “State coercion and worker solidarity in Italy (1915-1918): the moral and political content of social unrest”, pp. 41
44. Carlo Alberto Magni [1989] “Reputazione e credibilità di una moneta in un governo”, pp. 56
118. Massimo Baldini [1995] “Aggregation Factors and Aggregation Bias in Consumer Demand” pp. 33
134. Carlo Alberto Magni [1996] “Repeatability and a random real options a dynamic programming approach” pp. 23
137. David Avra Lane, Irene Poli, Michele Lalla, Alberto Roversato [1996] “Lezioni di probabilità e inferenza statistica - Esercizi svolti” pp. 302
139. Luisa Malaguti e Costanza Torricelli [1996] “Monetary policy and the term structure of interest rates”. pp. 30
162. David Lane [1996] “Is what is good for each best for all? Learning from others in the information contagion model” pp. 18