WP. 4
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (i): Disclosure risk assessment

Sampling as a way to reduce risk and create a Public Use File maintaining weighted totals,

Prepared by C. Casciano, D. Ichim, L. Corallo, ISTAT, Italy

Sampling as a way to reduce risk and create a Public Use File maintaining weighted totals

C. Casciano, L.Corallo, D. Ichim
Istituto Nazionale di Statistica, Via Cesare Balbo, 16, 00184 Rome, Italy
e-mail: {casciano, corallo, ichim}@istat.it

Abstract
Istat has initiated a research project addressing the problem of multiple releases from a single survey. In this paper some preliminary findings concerning the simultaneous release of a microdata file for research (MFR) purposes and a public use file (PUF) are presented. A brief analysis of the relationship between a MFR and a PUF is provided. A procedure aiming at the harmonisation of both disclosure risk and data utility is sketched. From an operational point of view, subsampling is the main statistical disclosure limitation technique used for deriving the PUF from the corresponding MFR. Different subsampling strategies that might be used in the SDC framework are illustrated. In this preliminary development stage, data from the Istat survey on Careers of Doctorate Holders is used.

Keywords: multiple microdata release, comparability, coherence, disclosure risk, data utility, subsampling, multivariate multi-domain allocation, balanced samples

1. Introduction

The release of statistical information in an international setting needs to find ways to reconcile adaptation to the national context and the need for global efficiency. The topic of harmonisation in statistical confidentiality limitation is crucial when facing the release of European statistics. Lately many structural changes have completely modified the way to deal with statistical disclosure limitation at European level. Many countries have to face more open attitudes toward the release of microdata; the upcoming revision of regulation 831/2002 will further improve changes in this sense. Member states have to find ways to balance national constraints and European needs. Among other objectives, the Essnet project on common tools and harmonised methodology for SDC in the ESS intends to test a possible strategy proposed by Ichim and Franconi (2010) to overcome difficulties stemming from the application of single rigid methodology for all member states.

Harmonisation of statistical disclosure control process is achieved through the harmonisation of methodology in the input phase and the provision of harmonised and objective measures for the output enabling de facto some flexibility and allowing for better adaptation to national context and better global efficiency. On the output phase, a set of benchmarking statistics might be defined for the type of data under analysis together with the corresponding quality criteria/thresholds. These are used to measure data utility and to put in practice the comparability concept.

The European Statistical Law 332/2009 mentions the opportunity to release both public use files (PUF) and microdata files for research purposes (MFR). This could be the main stimulus for investigating another harmonisation dimension; the multiple releases from a single survey should be harmonised from both disclosure risk and data utility point of view. This is a very complex task that could be fulfilled only by carefully analyzing the relationship between the two microdata releases.

The objective of this work is to propose a methodology for the release of a PUF derived from a MFR. The main tool of the already proposed comparability concept, i.e. the benchmarking statistics, is applied to develop a subsampling procedure aiming at the simultaneous release of two microdata files. Indeed, by a-priori defining the quality criteria to be maintained, it is illustrated how to derive a subsampling strategy guaranteeing the disclosure risk reduction below an acceptable threshold.

In section 2 a brief data description is given. In section 3, several aspects of the relationship between an MFR and a PUF are discussed. In section 4, the proposed subsampling methodology for the release of a PUF is detailed. To compare different implementation strategies, some preliminary results obtained when applying these strategies to the Careers of Doctorate Holders survey are described. Finally, in section 5, several conclusions and ideas for further testing are presented.

2. Data description

In 2009 Istat carried out for the first time the Career of Doctorate Holders (CDH) survey. This survey belongs to the system of surveys aiming at the characterisation of the education-occupation transition. CDH registers information on the occupational status of the PhD holders at 3 and 5 years from the date of completion of the PhD. Even if it was designed to be a census survey, only a response rate around 72% was observed. Thus, instead of 18500, only 12964 doctorate holders were interviewed. To adjust for non-response errors, an estimation procedure was adopted. The weights were computed using a calibration procedure, constraining on the known marginal distributions of citizenship (2 categories), PhD scientific area (14 categories) by gender and region.

The main focus of the CDH survey is the characterisation of the occupational status of the PhD holders. Aspects like facilities/difficulties to find a job (labour market entry), type of contract, earnings, type of work, job satisfaction, usefulness of the PhD for obtaining a job, etc are observed by the survey. Some statistics are already published in Brait and Strozza (2010). Weighted totals are generally published by PhD scientific area, by gender and by region (location of the university).

Besides the CDH data, structural information on the entire population of PhD holders is available, as transmitted by the universities. The release of samples from a data archive containing information on the entire population has two special advantages. Firstly, the opportunity to release samples with pre-defined disclosure risk should be stressed. Indeed, by controlling the sample design parameters, an *ex-ante* disclosure limitation method might be developed. Secondly, the availability of information on the entire population allows measuring *ex-post* the disclosure risk in terms of sample and population frequencies, without involving complex statistical models.

3. Multiple microdata release: MFR and PUF

In this section a brief discussion on some aspects related to the simultaneous release of a MFR and a PUF is given. More details might be found in Ichim et al (2011).

In this Istat project it is proposed to adopt an unique production process for both MFR and PUF. The existence of such process implies important efficiency gains from several points of view: a) the disclosure risk, b) coherence of the informative content of the files to be released, c) the physical creation of the microdata files and d) coherence of the associated metadata.

The disclosure risk assessment and management of a single microdata release are obviously related and constrained by those of other microdata releases. The same statement is valid also for the information content of the disseminated microdata files. Indeed, in the risk management frame work, when dealing with multiple microdata releases, different levels of access are created. When a less restrictive licence applies, i.e. a PUF licence, the disclosure risk should be more reduced since it is supposed that more users would have access. As the number and type of users increases, the adopted disclosure scenarios should take into account the existence of more external information.

The main assumption of this proposal is the existence of an unique hierarchical structure of the users. Since the MFR users already have access to more detailed and accurate information, it is supposed that the MFR users would not require the corresponding PUF. Only a sequential definition and production of the two microdata files would allow the data provider to take into account the users hierarchical structure. Moreover, if an unique production process applies, different coherent disclosure scenarios may be considered.

From the information content point of view, the usage of nested classifications should be recommended. Otherwise, in presence of non-nested hierarchical classifications, some differencing scenarios might be enabled. The level of detail associated to a less restrictive licence of use should be at least inferior to the one associated to a more restrictive licence.

When the two releases, the MFR and PUF, share the same structure and the information content shows a nested hierarchical structure, the PUF would even increase the MFR usability. For example, from the researcher point of view, the PUF would reduce the costs of using a more restrictive licence. Indeed, while waiting to receive

the required MFR, the researcher could start, on the basis of the information included in PUF, to organize the scientific work to be performed.

3.1 Contribution to the diffusion of the statistical literacy

It is our opinion that PUFs offering no guarantees in terms of data utility/data quality would not be very well accepted by the users. When the PUF satisfies some predefined quality standards, it would positively contribute to the diffusion of the statistical literacy. For example, PUFs would give students the opportunity to apply theoretical statistical knowledge in a real context.

The value added by a PUF is a straightforward consequence of its quality standards defined as the ability to simulate real applications. The file dimension, expressed as number of records and number of variables, provides a first quality indicator. Moreover, since the data production process and data quality are not extensively discussed in statistical lectures, any PUF could contribute to the reduction of this gap. At the same time, a large number of variables would favour the development of a critical reasoning on the variables meaning, their operational definition, the surveyed phenomenon, etc. The precision and accuracy of the estimates that could be derived using the PUF would significantly improve the conceptual learning. The relevance and timeliness should be other crucial aspects when deciding which PUF to produce.

4. Development of a procedure for the release of a public use microdata file

In this work it is proposed to derive the PUF from MFR by subsampling. Some records might be further modified if deemed necessary in some particular disclosure scenarios. Otherwise stated, it is proposed to release both PUF and MFR with the same level of detail in the hierarchical variables. Obviously, for the PUF minor precision and accuracy would be attained than for the MFR. The minor precision is a direct consequence of subsampling while the minor accuracy is a consequence of the fact than only few predefined estimates could be used as data utility constrains in our procedure. Of course, the global risk associated to a PUF should be smaller than the disclosure risk of a MFR. In this paper it is assumed that the same SDC identifying variables are used for both PUF and MFR. In this section a two steps procedure is described. First it will be shown how to take into account both disclosure risk and some data utility requirements when determining an optimal allocation. The second step of the procedure consists in drawing a random balanced sample, thus aiming at the approximate preservation of some weighted totals.

The procedure developed in this work aims at releasing a PUF maintaining some quality indicators and, simultaneously, at the preservation of the advantages of dealing with random samples; the PUF microdata should also be representative for the entire population, as the MFR does. Secondly, coherence with already published information should be assured. For example, the equality between the published totals, the ones derived from the MFR and the ones derived from the PUF should be aimed. Just to mention two obvious advantages of this restriction, this latter quality condition/indicator would increase the trust in the dissemination strategies of NSIs. Moreover this coherence between estimates would also contribute to disabling disclosure scenarios based on differencing. In principal it would be much easier to achieve this coherence between published totals when the PUF and MFR show a hierarchical structure. Instead of totals, some other descriptive statistics might be used as well. Only published totals are dealt with since they are one of the most important statistical products and they are among the first statistics to be computed.

The standard MFR dissemination model includes the definition of a disclosure scenario, risk assessment and protection with respect to the adopted disclosure scenario and with respect to some data utility requirements. Given some utility constraints, there are two alternatives for stating the dissemination problem of a PUF derived from a MFR. The first one, named strategy A, consists in the maximisation of the number of released safe units. The second one, named strategy B, consists in the minimisation of the released units at risk of disclosure.

At a first glance, it might be believed that strategy A is very easy and efficient. Indeed, since the risk of disclosure is supposed to be already estimated, the records could be easily classified in units at risk and safe units. In such situation, the release of only safe units might be very tempting. There would be an immense

efficiency gain since little additional efforts would be required. Only the updating of the calibration weights would be necessary. Anyway, there are several aspects that should be further investigated. First of all, if no further protection method would be applied, it would be implicitly assumed that the MFR and PUF share exactly the same disclosure scenario, i.e. identifying variables, means, intentions, tools etc. In practical situations, the validity of this assumption should be very carefully analysed.
Secondly, since all records at risk would be suppressed from the PUF, there might be some empty planned estimation domains. Consequently, it would be impossible to preserve the coherence with the already published totals. When each SDC identifying variable is among the variables defining the estimation domains, the existence of empty planned estimation domains cannot be ignored. Depending on data, this drawback might arise also in other settings.
Finally, since the disclosure risk is not a random variable, the release of safe units would not provide a random sample. A random sample from the safe units could be drawn, but such sample would represent only the "population" of safe units.

In strategy A, the disclosure risk is completely controlled and there is no restriction on data utility. The strategy A might be classified within the risk avoidance framework. Instead, strategy B proves to be more flexible. It mimics better the risk management framework. Indeed, by accepting that there might be some risk of disclosure in the PUF, data utility could also be increased.

A procedure based on sound statistical methods is presented in the remaining part of this section. The goal is to provide a PUF satisfying as many risk and utility requirements as possible. The imposed utility constraints include the preservation of the internal consistency of the records, the preservation of some totals, the reduction of the disclosure risk and the randomness of the sample.

It is proposed to relax the restrictions on the quality of some totals. The approach still aims at the preservation of some indicated totals; the required relaxation would be implemented in terms of precision/accuracy. The advantages derived from the designing a PUF and a MFR following the same hierarchical structure and sharing the same detail level of the hierarchical variables were already discussed. To reduce the risk, a statistical disclosure limitation (SDL) method should be applied. The constraint on the hierarchical structure eliminate recoding from the candidate list of SDL methods which could be applied. The application of other SDL techniques most used when disseminating microdata stemming from social surveys, e.g. suppression, PRAM or swapping might generate a non negligible utility loss. For example, PRAM or swapping should be modified to maintain the internal consistency of the records. Of course, modified versions of PRAM or swapping could be implemented, but their implementation would be very much application dependent. Consequently, the generalisation of their modified versions would be more complex without additional efforts.

Our proposal is based on another well-known SDL method, namely subsampling. This technique guarantees, by definition, the preservation of the hierarchical structure of the PUF with respect to the MFR, the preservation of the hierarchical detail and the preservation of the internal consistency of the records. When a random subsample is drawn, the randomness feature of the microdata file is obviously maintained, too.

Subsampling reduces the risk of disclosure by adding some uncertainty on the number of population units sharing the same modalities of the identifying variables. Hence, a possible intruder would have state that a sample unique is a population unique with an increased uncertainty.

Subsampling, does not guarantee, by default, neither a controlled reduction of the disclosure risk nor the preservation of some data utility indicators. Some adaptations are necessary in order to improve the risk-utility performance of this SDL technique.

Let us suppose that the MFR, here the considered population, has N records and M variables, $V_1,\ldots,V_M$. Suppose that from the previous analyses on MFR, a risk-related variable, R_1, is available, too. R_1 is a dichotomous variable indicating whether a unit is at risk of disclosure or not. Issues related to the derivation of R_1 are not further discussed.

It is assumed that the data-utility constraints may be expressed in terms of population totals. It is supposed that some MFR totals should be approximately preserved. In the specification phase, it should be mentioned which is the information to be preserved and on which estimation domains. Suppose that the information to be

preserved may be expressed in terms of variables $\{Y_1,\ldots,Y_K\}\subset\{V_1,\ldots,V_M\}$ and that the estimation domains may be expressed in terms of variables $\{D_1,\ldots,D_E\}\subset\{V_1,\ldots,V_M\}$. $\{D_1,\ldots,D_E\}$ could also be defined as cross-classifications of some variables among $V_1,\ldots,V_M$. Both $Y_1,\ldots,Y_K$ and $D_1,\ldots,D_E$ should be identified only after the consultation with survey and subject-matter experts.

There might be some overlap between the SDC identifying variables and the variables $\{D_1,\ldots,D_E\}$. Indeed, both types of variables express some structural information and the accessibility of external databases containing structural information cannot be ignored. As an example from the social surveys framework, variables like gender, age and marital status are generally considered SDC identifying variables. At the same time, the published totals generally refer to domains defined by (cross-classifications) of such variables, e.g number of employed persons by gender and age classes.

On the contrary, generally speaking, there is little chance to observe any overlap between the SDC key variables and $Y_1,\ldots,Y_K$. In principle, the SDC key variables are structural information available in external databases while the information-content variables $Y_1,\ldots,Y_K$ represent the aim of the survey, i.e. "collect information on topic … not available elsewhere".

The estimation variables tested in this work are Occupational status (Occup), Job search (JobS), Type of contract (Contract), Type of Work (Work), Income. Except for the last variable which is a continuous one, the others were dichotomized. The domain variables used in this application were Year of Completion of the Doctorate (Year, 2 modalities), Gender, Region of University (Region, 20 modalities) and Scientific Area (Area, 14 modalities).

The proposed procedure works as follows. First it determines how many units to draw in each domain and then it draws a fixed size sample.

4.1 Optimal allocation in stratified sampling

A goal of the stratified sampling is to increase the precision (reduce the variance) of estimates of population parameters inferred from a sample. All other things being equal, increased homogeneity of the population being sampled works to increase precision. By dividing the population of interest into non-overlapping subpopulations (sampling strata) that are more nearly homogeneous, selecting independent samples from each stratum, and combining estimates from the strata, more precise estimates than by directly sampling from the population can be computed.

Once the stratification has been defined, the question is how many sample units to allocate to each stratum. In the univariate case, the minimization of the cost design needed to achieve a target variance has a well-known solution:

$n_h = n\frac{N_h S_h/\sqrt{c_h}}{\sum(N_h S_h/\sqrt{c_h})}$, where n_h is the number of sample units allocated to stratum h, N_h is the number of population units in stratum h, c_h is the cost per unit in stratum h, S_h is the population standard deviation for the variable of interest in stratum h, and n is the total sample size. (S_h is usually estimated from frame information or earlier samples.)

In our SDC-PUF-MFR framework, precise estimates of several not highly correlated variables are needed. It is desirable to apply a method to find a good compromise allocation giving adequate precision for all of the variables of interest. A number of approaches have been used to find the optimal allocation of multiple variables on multiple domains. Bethel's approach is based on an acceptable coefficient of variation (CV) for each of the allocation variables. These CVs become constraints on a cost function to be minimized; then the following convex programming problem is solved:

$f(n_h)=\sum_{h=1}^{H} c_h n_h = \min$

$\mathrm{Var}({}_p\tilde{\bar{Y}}_{j_d})=\sum_{h=1}^{H_{j_d}}\frac{N_h^2}{n_h}\,{}_pS_h^2-\sum_{h=1}^{H_{j_d}}N_h\,{}_pS_h^2\le {}_pV_{j_d}^{*}$, …p=1,…,P, d=1,…,D, j_d=1,…, J_d, where d is a type of domain of interest, (d =1,..,D); j_d is the generic domain of interest of type d ($j_d=1,\ldots,J_d$, being J_d the number of the

domains of type d); H_{jd} is the number of strata containing the domain of interest j_d, $\mathrm{Var}({}_p\tilde{Y}_{j_d})$ is CV of the estimate for the p^{th} variable and ${}_pV^*_{j_d}$ is a prefixed planned level of the CV for the estimate of the p^{th} variable.

Thus, the problem of interest may be stated as, "Find a stratification and allocations to those strata minimizing the budget necessary to achieve predetermined maximum allowable coefficients of variation for two or more selected variables of interest."

For our purposes, Bethel's algorithm has been used in order to determine the optimal strata sizes in terms of reduction of the overall risk (cost function), keeping the CV level of the estimates below a 5% threshold for three combinations of the allocation and domain variables, see Table 1.

Table 1. Different variable combinations used for the application of the Bethel algorithm.

Combination	Domain variables	Allocation variables
1	Year, Gender, Area, Region	Occup, JobS, Contract, Income
2	Year, Gender, Area, Region	Occup, JobS, Contract, Work, Income
3	Year, Area	Occup, Contract

For each combination, six settings of the Bethel algorithm were tested, corresponding to different choices of the parameters. The six Bethel settings are described in table 2. The strata containing less than 2 units have always been considered as "take all" strata. Moreover, the strata containing no units at risk of disclosure might be subject to a total survey, too. When deciding to consider the risk variable among the stratification variables, the number of units at risk to be included in the allocation is somehow minimized. This might be seen as an obviuous consequence of the Bethel's algorithm philosophy, i.e. minimize the number of allocated units per stratum. It should be noted that the risk variable R1 was not used as domain variable since the latter is related to data utility. The final option in this work is related to the consideration of the total risk per stratum as cost function in the Bethel's algorithm.

Table 2. Description of different Bethel settings.

Setting	Risk.cost	Risk.strat	Cens.no.risk	Description
1	N	Y	N	The cost c_h does not depend on the risk R1. The risk R1 is used as stratification variable. The strata without units at risk are not considered "take all" strata.
2	N	Y	Y	The cost c_h does not depend on the risk R1. The risk variable R1 is used as stratification variable. The strata without units at risk are considered "take all" strata.
3	Y	Y	N	The cost c_h equals the sum of R1 by stratum. The risk R1 is used as stratification variable. The strata without units at risk are not considered "take all" strata.
4	Y	Y	Y	The cost c_h equals the sum of R1 by stratum. The risk R1 is used as stratification variable. The strata without units at risk are considered "take all" strata.
5	Y	N	N	The cost c_h equals the sum of R1 by stratum. The risk R1 is not used as stratification variable. The strata without units at risk are not considered "take all" strata.
6	Y	N	Y	The cost c_h equals the sum by stratum of R1. The risk R1 is not used as stratification variable. The strata without units at risk are considered "take all" strata.

We want to select samples of fixed size from a population of respondents to the CDH survey, balanced on socio-demographic variables such as *Gender*, *Year of Doctorate Completion*, *Scientific area*, *Region*. Each sample has been selected under a stratified random sampling design, whose optimal strata size has been determined via Bethel algorithm.

The sampling sizes obtained using the different combinations (of allocation and domain variables) and different Bethel settings are presented in Table 3. The column Risk.cost indicates whether the risk was used as the minimization cost of the Bethel algorithm. The column Risk.strata indicates whether the variable risk was used as a stratification variable. The column Cens.no.risk indicates whether the strata without any unit at risk were "take all" strata. The column #Strata indicates the number of strata in each allocation setting. The column #Cens.strata indicates the number of "take all" strata. The column #Cens.units indicates the number of units in the "take all" strata, when this option was used. The columns Size.Bethel, Size.Prop and Size.Equal show the total number of allocated units when applying the Bethel algorithm, the proportional to size and the equal allocations, respectively. The gains in terms of sample size reduction obtained w.r.t. the proportional and equal allocations may be assessed by comparing these three rows. It may be observed that, even if the differences between the three allocations are not very striking, the usage of the Bethel algorithm always provides the smallest sampling fraction. Consequently, it is reasonable to expect the drawing of the sample with smaller number of units at risk of disclosure, while preserving the same data quality standards expressed as precision of the estimates. Moreover, since the Bethel allocation is optimized w.r.t. some allocation variable while the other

two allocations are not optimized, the distribution of the resulting samples is generally quite different among the estimation domains. The columns Max.Bethel-Prop and Max.Bethel-Equal indicate the maximum by strata of the absolute differences between the Bethel and proportional and equal allocations. It may be observed that even when the three allocations are almost equal, there are some strata which might be subject to very dissimilar allocations. For example, in the setting called 3.3, highlighted in gray, the Bethel, proportional and equal allocations procedures provide around 8800 units. But there is a strata where the difference between the Bethel and proportional allocation is equal to 189 units and there is a strata where the difference between the Bethel and equal allocation is 389.

Once the optimal allocation has been determined, a sample should be drawn. Using the Bethel algorithm, the optimal allocation is derived in order to take into account some predefined precision requirements. In the next phase, the selection phase, a sample should be drawn in such a way that it fulfills some predefined accuracy requirements. In the next section a methodology enabling the control of the accuracy constraints is illustrated.

4.2 Balanced sampling

Balanced sampling consists in drawing random samples which provide exact estimates for some auxiliary variables. The Cube method (see Deville and Tillé, 2004, 2005) enables the selection of balanced samples, with unequal probabilities and a non-restricted number of balancing variables.

A sampling design *s* is said to be balanced on the auxiliary variables $\mathbf{x} = (x_1 \ldots x_j \ldots x_p)$ if and only if the balancing equations given by $\hat{\mathbf{x}}_\pi = \mathbf{x}$ are satisfied, where $\mathbf{x}$ is the vector of known population totals, $\hat{\mathbf{x}}_\pi$ is the Horvitz-Thompson estimator expressed as $\hat{\mathbf{x}}_\pi = \sum_{k \in s} \mathbf{x}_k / \pi_k$, $\pi_k = \Pr(k \in s)$ being the first order inclusion probability of unit *k*.

A sampling design balanced on the variable $x_k = \pi_k$ is of fixed size, as $\sum_{k \in s} \frac{x_k}{\pi_k} = \sum_{k \in s} 1 = n$. A sampling design balanced on the variable $x_k = 1$ ensures that the population size *N* is perfectly estimated, as $\sum_{k \in s} \frac{x_k}{\pi_k} = \sum_{k \in s} \frac{1}{\pi_k} = \hat{N} = \sum_{k \in U} x_k = \sum_{k \in U} 1 = N$, where *U* denotes the entire population. Note that these two variables are always available in the population. When the variable of interest which is well explained by the balancing variables, a balanced sampling design generates large accuracy gains w.r.t. afixed size unequal probability sampling design.

In case of a stratified random sampling design, the sampling design is balanced by strata on the variable *x* if $\sum_{k \in U_h} \frac{S_k x_k}{\pi_k} = \sum_{k \in U_h} x_k$ for all $h = 1, \ldots, H$, where S_k is the random variable indicating the inclusion of the unit k into the sample.

Stratified balanced sampling can be performed by selecting a sample directly from the whole population. The condition of balance by strata is equivalent to the expression:

$\sum_{k \in U} \frac{S_k \left(x_k \, 1_{k \in U_h} \right)}{\pi_k} = \sum_{k \in U} x_k \, 1_{k \in U_h}$ for all $h = 1, \ldots, H$.

We thus only need to select a sample in *U*, balanced on the variables equal to the product of the balancing variables $x_1 \ldots x_j \ldots x_p$ and the indicator variables: $\mathbf{1}_{k \in U_h} = \begin{cases} 1 & k \in U_h \\ 0 & k \notin U_h \end{cases}$ which means balancing on the $H \times p$ variables. This method has the main drawbacks, that if is too big we cannot search for a sample that causes a small difference to the balancing state, because the number of possible samples is too large; besides that, all strata do not have the same balancing quality.

The cube method provides a general solution to the problem of selecting balanced samples, with any vector of inclusion probabilities and a certain number of balancing variables.The CUBE method consists in two steps, named flight and landing phase, see Tillè (2006). During the flight phase, if all the constraints, i.e. the balancing equations, are exactly satisfied; the algorithm stops as soon as it determines a perfectly balanced sample.

Otherwise it stops when the convergence to a balance solution is not achieved1; in this case, for a subset of units the inclusion in the sample is still uncertain. Then the landing phase starts; it searchs randomly a sample which achieves the best approximation to the balancing equations.

The landing phase implies a weakening of some constraints, according to three possible criteria: a) after listing some priorities, the balancing variables are progressively abandoned; b) the landing phase is performed by considering all the possible samples among the uncertain units, and selecting those providing a low difference to the balance (the difference to the balance expressed by some distance measure); and c) the landing phase is performed as in b, but only considering the samples whose size equals the sum of inclusion probabilities; in this case the result is a fixed sample size.

Optimal overall and strata sample sizes provided by Bethel algorithm represent the vector of inclusion probabilities $\pi = [\pi_k]$ needed to apply the CUBE Macro.

In this preliminary testing phase of the Istat project, to ensure that the population size *N* and the optimal sample size *n* would be perfectly estimated, $x_k = 1$ and $x_k = \pi_k$ respectively have been introduced as balancing variables. In addition, three socio-demographic variables, i.e. *Gender*, *Year of Doctorate Completion*, *Scientific Area* have been separately considered in the system of balancing equations as stratification variables, for the same balancing variables, i.e. $x_k = 1$ and $x_k = \pi_k$. The known totals were the corresponding estimates derived by means of the survey weights. This means that the marginal frequency distributions by *Gender*, *Year of Doctorate Completion* and *Scientific Area* are exactly maintained together with the strata population and sample sizes indicated by the Bethel algorithm.

The third criterium in the landing phase was used; this cube option may be applied using a maximum number of 18 constraints to achieve a solution in a reasonable time. The balancing equations used in this testing phase are exactly 18.

It should be stressed that other data utility requirements are implicitly satified. Indeed, the c) convergence criterium of the cube method guarantees the minimisation of the variation of the input inclusion probabilities $\pi = [\pi_k]$. Consequently, the corresponding weights are not significantly adjusted; it follows that some other data utility requirements expressed as weighted totals are approximatively satisfied. In our application this statement holds because some data utility-related variables, e,g, *Occup*, were used as allocation/estimation variables in the Bethel's procedure.

For the five data-utility variables used in this application, namely *Occup, JobS*, *Contract*, *Work*, *Income*, estimated totals were compared to the ones derived from the original MFR. The estimated totals were computed for each stratum defined by *Gender, Year of Completion of the Doctorate*, *Scientific Area* and macro-region. In Figure 1, the median of the relative absolute errors are shown. The results obtained for each combination of of allocation and domain variables, see table 1, are highlighted by differently coloured rectangles. As expected, as the number of allocation/domain variables increases, the differences between the estimated totals and the MFR known totals decrease. Moreover, it may be observed that for variables with very skewed distributions like *Occup* and *Contract*, the computed differences between the estimated and known totals are relatively higher than those corresponding to other variables. This tendency is confirmed also by the means of absoluted relative errors.

In table 3, the correlations between the ratio of estimated totals and the sample size and the corresponding population quantities are illustrated in the columns called *Occup, JobS, Contract, Work*. For the sample called 3.4, two outlying values were cancelled. From table 3, it may be noted that the settings using a census in the strata without units at risk do not necessarily perform better from the data-utility point of view. The correlations between the estimated and known MFR means/totals are generally smaller when the census option is used than in those settings where the entire population is randomly sampled. Indeed, in this preliminary testing, it seems that it is more difficult to achive convergence to the balancing equations when a large number of units is not subject to a random selection.

In the column Risk of Table 3, the number of sampled units at risk of disclosure is indicated. As the number of variables used as domain variables in the Bethel's procedure increased, the number of strata increases and,

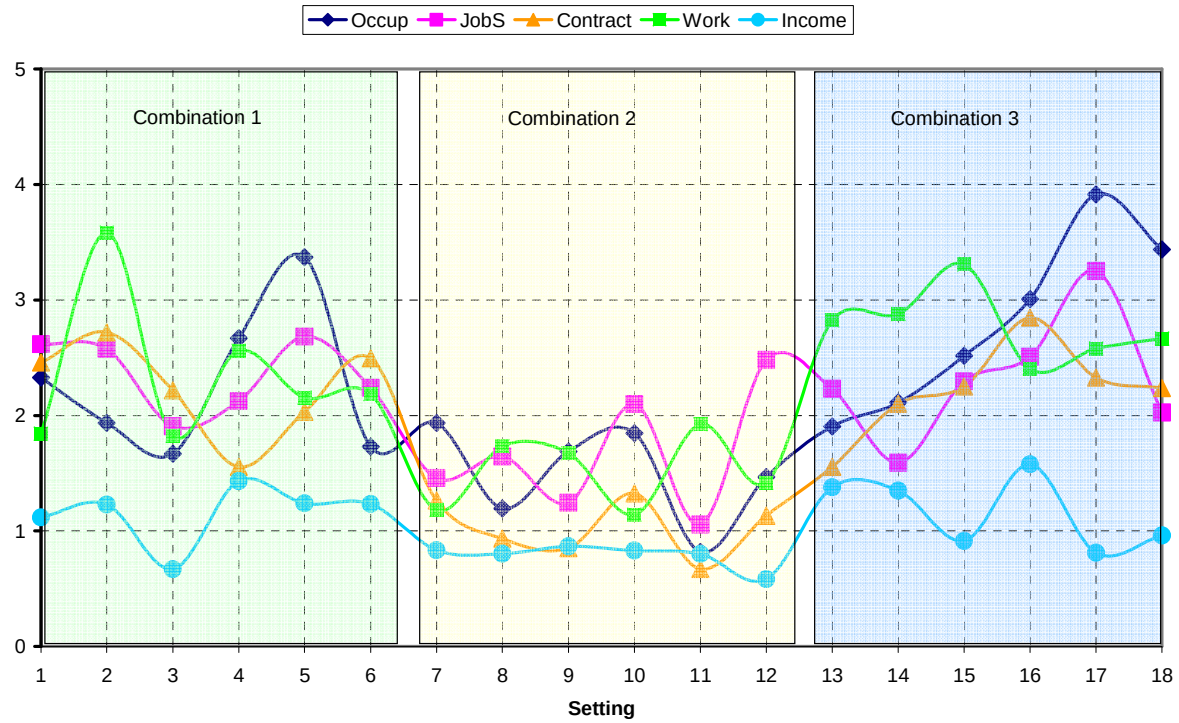
[1] Due to rounding problems, for example

consequently, the number of sampled units at risk increases. Generally, when the strata containing only safe units are "take all" strata, the number of sampled units at risk decreases, even if the reduction is not always significant.Among the 18 settings tested, the one highlighted in gray seems to achieve the best compromise between the disclosure risk reduction and the data-utility preservation. In this setting, when applying the Bethel's algorithm, the risk of disclosure was used as both cost and stratification variable.

Table 3. Risk and data utility (correlations) results.

C.S	Risk.cost	Risk.strat	Cens.no.risk	# Strata	#Cens.strata	#Cens.units	Size Bethel	Size Prop.	Size Equal	Max.Bethel-Prop	Max.Bethel-Equal	Risk	Occup	JobS	Contract	Work	Income
1.1	N	Y	N	925	153	252	4933	5391	5550	459	618	1366	0.88	0.97	0.97	0.99	0.99
1.2	N	Y	Y	925	214	704	5105	5547	5550	443	446	1333	0.92	0.99	0.94	0.97	0.99
1.3	Y	Y	N	925	204	558	5239	5719	5550	480	311	1335	0.92	0.98	0.95	0.99	0.99
1.4	Y	Y	Y	925	235	814	5330	5781	5550	451	220	1354	0.87	0.99	0.95	0.97	0.99
1.5	Y	N	N	925	240	687	5555	5953	6475	399	921	1490	0.86	0.98	0.97	0.98	0.98
1.6	Y	N	Y	925	269	983	5649	6094	6475	446	827	1525	0.91	0.98	0.95	0.97	0.99
2.1	N	Y	N	925	306	1614	8725	9256	9250	530	524	2194	0.83	0.91	0.99	0.97	1.00
2.2	N	Y	Y	925	352	1919	8827	9324	9250	498	424	2177	0.56	0.81	0.99	0.94	0.99
2.3	Y	Y	N	925	416	3229	8955	9424	9250	468	294	2149	0.78	0.91	0.99	0.91	1.00
2.4	Y	Y	Y	925	451	3398	9045	9511	9250	466	205	2163	0.64	0.88	0.97	0.95	0.99
2.5	Y	N	N	925	426	3243	9151	9601	9250	451	100	2232	0.63	0.87	0.99	0.86	1.00
2.6	Y	N	Y	925	457	3399	9222	9669	9250	446	84	2233	0.55	0.78	0.96	0.94	0.99
3.1	N	Y	N	56	0	0	4745	4773	4760	138	132	1272	0.96	0.99	0.92	0.96	0.98
3.2	N	Y	Y	56	28	9761	10320	10346	10360	166	630	559	0.52	0.79	0.41	0.83	0.98
3.3	Y	Y	N	56	21	5844	8812	8841	8848	189	389	564	0.77	0.94	0.93	0.97	0.99
3.4	Y	Y	Y	56	28	9761	10323	10349	10360	166	630	562	0.56*	0.84	0.59	0.88	0.99
3.5	Y	N	N	28	0	0	4760	4774	4788	176	88	1270	0.95	0.99	0.98	0.99	0.99
3.6	Y	N	Y	28	0	0	4759	4774	4788	176	88	1247	0.91	0.99	0.98	0.99	0.98

Figure 1. Median of absolute relative errors.



5. Conclusions and further work

In this paper a strategy for deriving a public use file from a microdata file for research purposes was described. The proposed methodology is based on the assumption that the PUF and MFR show a hierarchical structure from both point of view of disclosure risk and data utility. It was illustrated how to derive a PUF from an MFR by preserving some predefined quality levels. It was shown how to adapt subsampling in order to take into account the risk of disclosure and, at the same time, to guarantee that some predefined precision levels of some estimates could still be achieved. By decreasing the required precision levels, the disclosure risk could be decreased too. In the second step of the proposed procedure, it was discussed how to draw a sample which maintains some weighted totals. Here the preservation of some weighted totals was considered as the main data utility indicator. The usage of the balanced sampling in the SDC framework was illustrated.

The obtained results show that this strategy could be exploited in practical settings for the release of public use files. Anyway, the survey and subject-matter experts should clearly indicate which information should be preserved, i.e. which weighted totals.

There are several issues that will be further investigated. First, the relationship between coefficients of variation and disclosure risk will be explored, together with different options of including the risk of disclosure in the sampling design. Second, as regards the second step of the illustrated procedure, the other landing options of the cube method will be tested. We expect to introduce an utility-priority approach into the way to deal with the balancing equations, i.e. the weighted totals. Finally, the usage of other data utility constraints will be investigated.

References

Brait F, Strozza M. (2010), L'inserimento professionale dei dottori di ricerca. Available on www.istat.it.

Deville J., Tillé Y. (2004), Efficient Balanced Sampling: The Cube Method, Biometrika, Vol. 91, No. 4 (Dec., 2004) (pp. 893-912)

Deville J. Tillé Y. (2005), Variance approximation under balanced sampling,Journal of Statistical Planning and Inference, 128:569-591.

Ichim, D., Franconi, L. (2010), Strategies to Achieve SDC Harmonisation at European Level: Multiple Countries, Multiple Files, Multiple Surveys, J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, Lecture Notes in Computer Science 6344, Springer-Verlag Berlin Heidelberg, pp. 284–296.

Ichim D. Casciano C., Foschi F. Corallo C., Franconi F. (2011), Progettazione del rilascio di un file di dati elementari per scopi di ricerca scientifica ed il rilascio di un file di dati elementari ad uso pubblico. Indagine: Inserimento professionale dei dottori di ricerca. Istat internal report; (in italian)

Tillé Y. (2006), Sampling algorithms, Springer, New York.