WP. 36
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 (vii): Trans-border access to microdata

Farm Structure Survey: Considerations on the release of a European Microdata File for research purposes

Prepared by L. Franconi, D. Ichim, and L. Corallo, ISTAT, Italy

Farm Structure Survey: Considerations on the Release of a European Microdata File for Research Purposes

L. Franconi[(1)], D. Ichim[(1)], L. Corallo[(1)]

[(1)]Istituto Nazionale di Statistica, Via Cesare Balbo, 16, 00184 Rome, Italy e-mail: {franconi,ichim,corallo}@istat.it

Abstract

This document indicates the main issues to be taken into account in the development and implementation of a methodology for statistical disclosure control to be applied to the Farm Structure Survey (FSS) microdata for the creation of a European micro-data file for purposes of scientific research.To illustrate the procedures for disclosure risk assessment and the corresponding methods of disclosure limitation, data from the Italian FSS 2005 and Dutch FSS 2007 were used. This paper presents the main findings of a more detailed study commissioned by Eurostat.

1. Introduction: stages for the development of a micro-data file for research

When releasing a micro-data file, the confidentiality of the respondents need to be preserved. For this reason disclosure risk is analysed. If such risk is consider too high for the type of users, by reducing the information content of the micro-data file, the risk of disclosure is reduced. However, the users need micro-data that resemble as much as possible to the original micro-data file The aim of the development of any micro-data file for external release is to find the right balance between data confidentiality and data utility.

The micro-data file we consider in this document should be released exclusively for scientific research purposes. Consequently, a rigorous study of possible disclosure scenarios, including spontaneous identification scenarios, is carried out. Then a careful risk assessment analysis is performed. The protection is achieved using variable suppression, variable aggregation and perturbation. Data utility constraints is taken into account when the perturbation method is applied.

In Section 2 the main survey characteristics are described and some data analysis is performed in order to highlight the main features that need to be taken into account when proposing a data protection method. The statistical disclosure control methodology is presented in Section 3. Details on the disclosure risk evaluation, disclosure limitation and data utility are also given in Section 4. Section 5 outlines methods to be used for monitoring the impact of disclosure limitation on information loss. The conclusions are presented in Section 6.

2. Description of the European Farm Structure Survey (FSS) and its characteristics

Although the Farm Structure Survey (FSS) is organised in all Member States of the European Union on a harmonised legal framework (http://epp.eurostat.ec.europa.eu/portal/page/portal/agriculture/legislation), national peculiarities determine different structures both in type of survey and type of data.

The FSS collects information on the agricultural holdings (the survey unit by definition) in all Member States at different geographical levels (Member States, regions, districts) and over periods (follow up of the changes in the agricultural sector). Thus the FSS provides a base for decision making in agricultural frameworks. Whereas the characteristics are based on community legislation, the same information is available for all countries for each survey. From the very beginning it should be noted that the survey collects very little information on the economic performance of the agricultural holdings. The main structural agricultural variables may be classified in four groups: (i) general overview (key variables), (ii) detailed data on land use, (iii) detailed data on livestock and (iv) detailed data on special interest topics: farm labour force, rural development issues as well as management and practices.

2.1 Survey characteristics and implication on SDC methods

Survey Strategies

Member States (MS) collect FSS data using both censuses and sample surveys. All MS conduct a survey at least every 10 years, but 7 of them (BE, LU, NL, FI, SE, UK and NO) conduct a census each survey round. In some countries like UK and Norway a mixed data collection mode is really used (sample survey and census). This feature should be seriously investigated. From the micro-data release point of view, in case of a census micro-data file, a sampling should be really applied. Instead, this kind of solution (sub-sampling), should be carefully analyzed in case the original data is already a sample in order to evaluate the impact of a double sampling scheme on statistical properties of the released file. Moreover, the mixed data collection modes could further complicate the sub-sampling design.

The statistical unit is always the agricultural holding. For all MS, the target population is defined by those agricultural holdings exceeding a certain national threshold.

Sampling Design

Generally, the frame is updated in order to account for the merging and separation phenomena. In case of sampling surveys, a stratified random sample design is usually adopted by the MS. As common for establishment surveys, a census is performed in the strata containing the largest farms. The strata are derived as cross-classification of variables expressing dimension (UAA, utilised agricultural area; LSU, livestock

unit ;ESU, total standard gross margin of the holding), geographical location and typology. As mentioned in the regulations governing the FSS, the samples are taken out in order to guarantee some predefined accuracy levels on some variables like UAA and ESU.

Response Rate and Extrapolation factor

A very high response rate might obviously increase the disclosure risk, especially for the largest enterprises. The survey in countries like Italy, Norway, Romania, Hungary, Greece, France and Cyprus presents very low non-response rates, e.g. 10%, 10%, 4%, 0.4% 1.5%, 0.1%, and 7.5% respectively.

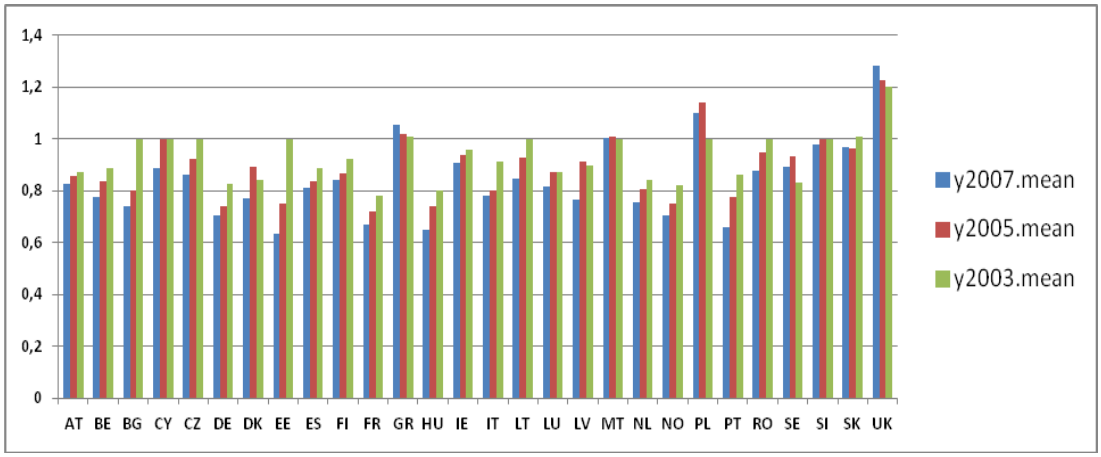
The design weights are generally updated to account for unit non-response.

The final weight of each unit is computed as a product of three factors: sampling weight, total non-response and calibration adjustment factors. The latter adjustments were performed in order to preserve some population characteristics, as derived from the agricultural census.

2.2 FSS micro-data main features

In this section, by analysing the number of holdings in the MSs at different geographical levels and time periods, it is observed that no significant change may be observed between different reference years, (2000-2007), see **Fig1**. Consequently, it should be sufficient to develop the statistical disclosure methodology focussing on a single reference year.

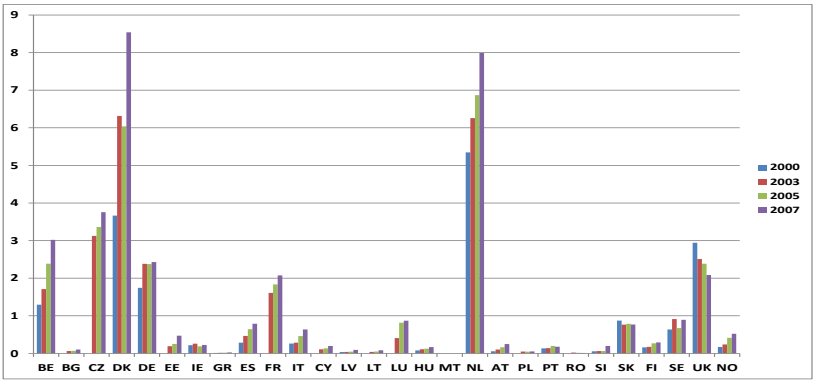
Fig 1: Relative variations of the mean number of holdings with respect to year 2000 at NUTS2 level.



Moreover, exploring the similarity between a region of a large country and a whole small country, it is not an easy task to find a discriminant dimension between large and small countries. This means that the disclosure risk evaluation could hardly be performed by analysing the agricultural phenomenon in a single MS.

Fig 2 shows that countries with small dimension in terms of land area appear to be closer to the x axis, i.e., generally speaking, "small" countries do not have a significant number of large holdings over the total number of holdings. Another feature present in Fig 2 is the fact that the percentage of large holdings continuously increases while the percentage of small holdings continuously decreases over time. This information is relevant from the risk of disclosure point of view since large holdings always shows a greater risk of disclosure than small holdings.

Fig 2: Percentage of large holdings at NUTS0 level in each wave of FSS.



3. Disclosure Scenarios and Risk Analysis

To assess the disclosure risk, one should make realistic assumptions about what an intruder might know about respondents and what information would be available to him to match against the micro-data and potentially make an identification and a subsequent disclosure. These assumptions are known as disclosure scenarios, see Hundepool (2009). In this section some re-identification scenarious are described. A very basic and preliminary analysis was performed using the Italian FSS 2005 and the Dutch FSS 2007 data.

Micro-data files are released only after removing direct identifying variables. Other variables in the micro-data file could be used as indirect identifiers such as geographical location, farming type, standard gross margin, etc.. An identifying key variable is defined by compounding several identifying variables and can be used by an intruder for re-identification purposes.

As mentioned in the Commission Regulation 831/2002, the anonymized micro-data file would be released only for scientific research purposes. It should be assumed that the (academic) researchers are bona-fide users. Consequently, it is supposed that the researcher would not deliberately try to identify any farm. It follows that any (record)

linkage with an (external) register might not be deemed a realistic hypothesis. This is due to the huge amount of resources generally involved in any (record) linkage process.

The anonymized micro-data file would be released only to researchers signing an agreement with Eurostat and NSIs. It follows that the users could not generally be colleagues, and not even competitors, of the observed statistical units. Thus, the nosy colleague scenario could not be deemed a realistic disclosure scenario.

3.1 Spontaneous Identification Based on Structural Variables

Using the Italian FSS 2005 survey data, by a careful analysis of the observed variables, the following categorical structural variables might considered identifying:

(i) *Area status* (A05) – 3 categories

(ii) *SGM*[1] *region code* (A07) – NUTS2 - 21 categories for Italy (other "geographical "variables", e.g. A04A *Survey District NUTS code* or A04D *Municipality code for objective zones* could also be used)

(iii) *Gender* (L011) – 3 categories (the holding might be a legal person)

(iv) *Age group* (L012) – 7 categories

For the Italian FSS 2005, in 4% of the 649 combinations of these four variables the sample frequency is equal to 1 and, in many of these combinations the population frequency, is not greater than 2. In 3% of combinations of the above four variables the sample frequency equals to 2 and the corresponding population frequencies is not always so large. In about 2% of combinations, the population frequency corresponding to unique and double sample cases is less than 5. These latter units should/could be considered at risk of re-identification. For the Dutch FSS 2007 (a census), in 15% of the 27 combinations of these four variables the frequency is equal to 1 and in 11% of combinations the was equal to 2.

The variable *Utilised agricultural area* (A11) is a continuous variable and it could be used to identify a farm. Anyway, except for few units, the variable A11 has a compact range of values. With regard to the few extremely large values, see for example the Italian FSS2005 micro-data, it should be noted that they correspond an unique agricultural activity: they possess only *Permanent grassland and meadow(F)*. Since agricultural areas with such characteristics are obviously visible without any effort, it may be considered that the possible intruder wouldn't increase his

[1] **Standard Gross Margin** For each activity on a holding, or: farm, (e.g. wheat, dairy cow or vineyard), a standard gross margin (SGM) is estimated, based on the area (or the number of heads) and a regional coefficient. The sum of all margins, for all activities of a given farm, is referred to as the economic size of that farm. The economic size is expressed in European Size Units (ESU), 1 ESU being equal to 1200 Euro of SGM.

knowledge. Consequently these farms could not considered at risk of re-identification in this disclosure scenario.

As for the Legal personality, in Italy 96% of cases are concentrated in two of the 6 possible categories; taking into account the low frequency of the other 4 categories, this variable was not considered an identifying variable. Moreover, if the legal personality of the holding were used as an identifying variable, the variables gender and age group would significantly diminish their identification power.

3.2 Spontaneous Identification Scenario

As already stated, it is assumed that the researchers are bona-fide researchers. However, an individual re-identification might occur because of the particular characteristics of the survey. A researcher might (unintentionally) use some previous and very detailed knowledge for the re-identification of a farm.

Even if the economic dimension of the farms is not the main FSS objective, it should be noted that some of the farms are well-known to the general public, not only to the researchers. Moreover, as a census is generally conducted for the largest farms, it is known that the most (economically) important agricultural enterprises are included in the sample. The re-identification of such well-known enterprises could be performed using the *Standard Gross Margin*. Finally, it should be reminded that some MS systematically conduct an agricultural census.

Finally it should be taken into consideration the territorial aspect of the survey. Since any agricultural phenomenon is highly related to the territorial characteristics, the geographical location of the farms is essential. Consequently, the release of a micro-data file without a detailed geographical information would significantly reduce its research potential. Unfortunately, the territoriality has some consequences also on the risk of re-identification. For example, the dissemination/access of micro-data at NUTS3 level might not be deemed feasible by those MS having a very reduced number of holdings. Moreover, the number of large enterprises (in terms of ESU or total SGM or UAA) should also be considered. Such dissemination thresholds (the minimum number of holdings) should be defined by each MS, according to their own dissemination policy.

The (economic) activity of the farms is expressed by phenomena that are highly visible, e.g. crops and livestock. This visibility obviously favours the spontaneous identification. For example, a farm that is specialist in field crops, e.g. cereals, oilseeds, etc., might be identified by simply observing the field. In the statistical disclosure control framework, this means that some external information is readily available to anyone. Moreover, this information might be quite detailed, as it is registered also in the micro-data file. In conclusion, each farm might theoretically be at risk of re-identification in this spontaneous identification scenario.

4. Disclosure limitation analysis

The FSS statistical disclosure limitation procedure aims at managing the protection of agricultural holdings. The FSS microdata contains both categorical and continuous variables; such difference in type requires different treatments. SDC procedure needs a combination of several steps: the suppression of some identifying variables, the aggregation of some detailed categorical variable and the perturbation of some numerical variables.

4.1 Suppression of Some Identifying Categorical Variables

For each MS, the variables directly indicating the geographical location of a statistical unit in the FSS are:

1. A04D *Municipality code for objective zones 2000* (LAU1)
2. A04A *Survey District NUTS code* (In Italy this corresponds to NUTS3)
3. A07 *SGM region code* (In Italy this corresponds to NUTS2).
4. A05 *Area status*
5. A03 *Agricultural areas with environmental restrictions*

An analysis of the different impact of the geographical variables on possible identification of units has been carried out using as a case study the Italian FSS2005. We consider the set of identifying variables defined in Section 3.1 and report the number of unique and double combinations.

If the variable A04D was used to give information about the geographical location of the farm, 61% of the combinations of the categorical identifying variables would be sample unique cases. Instead if the variable A04A was used, 12% would be sample unique cases. In both cases, the number of unique cases could be considered too high. If the microdata file would be released at NUTS2 level - only 4% remain unique cases in this setting. For the Ducth FSS2007 data, 15% of the combinations of the categorical identifying variables are unique cases.

For the Italian FSS2005, the suppression of the variable A05 produces as a result that the sample frequencies of the combinations of A07, L011 and L012 are all greater than 3. However, variable A05 is related to the incentives an agricultural holding might receive due to the agricultural and climatic conditions it operates. This variable might be important for analysts. It follows that the suppression of A05 might be considered by some MS a significant data utility loss.

4.2 Analysis of the issues related to the transformation of too detailed variables

In order to reduce the detail of the released information, some variables might be aggregated. In this section, by variable aggregation we mean their addition. Of course, such method could be applied only to variables having the same measurement "unit" (hectares, number of animals, etc). In order to aggregate variables without losing too much information some issues need to be addressed.

At a first glance, the aggregation of variables could be linked to the percentages of zero values (missing phenomenon). Variables presenting a large percentage of zero values (missing phenomenon) could be more easily added up without a statistically significant impact on the agricultural characteristic represented. However, the analysis of zero values carried out on the Italian and Dutch data demonstrates the strong regional character of the agricultural phenomenon and, the sparsity of the data. Whatever disclosure limitation methodology will be applied it should preserve these two aspects.

Another feature that is essential when defining type of aggregations is related to the derivation of the SGM and farming type variables. Indeed, we see two possible strategies for the release of such essential variables:

i. release the total SGM and the farming type in the microdata file fo research (MFR) as they are available in the original micro-data (making a further analysis on the level of detail for the release of the farming type);
ii. recalculation of the values of the total SGM and the farming type based on the recoded and perturbed variables due to the anonymisation process.

The second option is definitively more cumbersome and labour intensive than the first one; indeed if some variables involved in the calculation are recoded or modified because of the protection procedure the correct SGM cannot be recovered. Moreover the SGM and farming type should be recalculated for each farm of each member state. The gain in releasing approximate SGM could be of little use.

Another issue to be considered when aggregating variables is the regional variability of their corresponding SGM coefficients. The information loss in the SGM due to the aggregation of some variables is proportional to (depends on) the standard coefficients difference between such variables. At this detailed level, the regional differences in coefficients are not generally significant. Even if this statement cannot be easily generalized, it is reasonable to assume that, for the majority of regions, the differences in standard coefficients between such detailed variables (belonging to the same group of variables) are not significant.

Finally, classification issues are crucial when aggregating FSS variables: in fact, the possibility to exactly classify each agricultural holding into one of the 70 farming type *subdivisions* or indeed the 50 *the particular types* depends also on the type of aggregation adopted. In order to preserve the various hierarchical level of the farming type classification an unavoidable step is the deep analysis of the constraints that are to be maintained in the aggregation process (see Eurostat, 2003).,

4.3 Possible Treatments of Continuous Variables

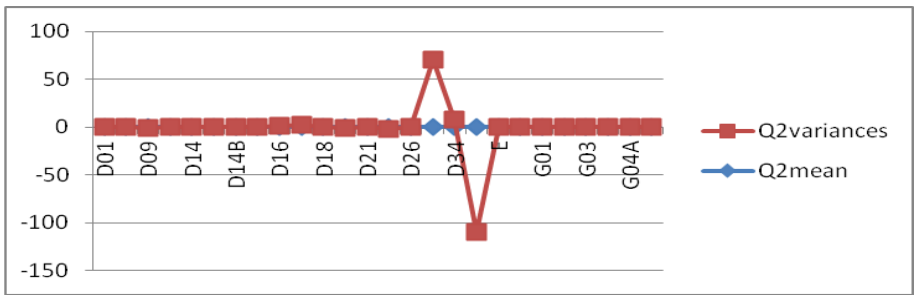
In a business framework, like the FSS, it is widely recognized that the continuous variables should receive special attention. Generally, the skew distributions of the continuous economic variables might allow a straightforward re-identification of the most dominant farms/enterprises. The most dominant farms are also the most visible

ones. Consequently, a perturbation of the continuous variables should be deemed necessary. In the SDC literature, there are many protection methods. When selecting a perturbation method, the possible usages of the micro-data file for research purposes should be considered, too. The FSS data is generally used for descriptive statistics. Hence a SDC method flexible enough to maintain such descriptive statistics should be considered.The individual ranking (IR) (Defays and Anward, 1998) is one of the simplest perturbation methods; it has been modified in order to preserve the weighted means (generating descriptive statistics) by using the extrapolation factor as weighting variable. Further criteria have been introduced when grouping the values to preserve, as much as possible, the meaning that zero values have in this survey. In fact, when averaging non-zero and zero values, a zero value (missing phenomenon) is replaced by a non-zero value (existing phenomenon). In order to control and reduce this drawback, when non-zero values were to be averaged with some zero values, the latter could be looked for within the farms sharing the same farming type with the units corresponding to the non-zero values. In this way, the artificial phenomenon would be created only within some farms having similar agricultural characteristics with the original farm. The individual ranking was applied using as microaggregation parameter *k=3*, and *SGM region code* as blocking variable.

5. Information loss assessment

When applying whatever statistical disclosure limitation methodology, some information loss is unavoidable When applied as proposed in Section 5, the IR would preserve, for each numerical variable, the NUTS2 weighted means. This should be considered the most important data utility constraint because these weighted means are the most used statistical tools. For each numerical variable and for each combination of NUTS2 and *general types* of farming, the weighted means/variances were computed both for the original and perturbed data. Then, the percentage variation was computed The skew distributions observed are a consequence of the sparsity of the represented phenomenon. The weighted means by region and farming type (a deeper detail than the regional level), were slightly modified; the most significant modifications appear to be present in correspondence of variables with a high number of missing values (see **Fig 3**).

Fig 3 the second quartile Q2 the variances variations over the combinations of A07 and A06



It should also be noticed that the information loss would be surely different if the Individual ranking would be applied at NUTS3 level. The release of NUTS3 geographical level together with a "heavier" IR or release NUTS2 geographical level together with a "light" perturbation are all possibilities that need to be further investigated.

6 Conclusion

The process of developing a European microdata file for research from the FSS is complex due to the peculiar nature of the data and the heterogeneity of the collection modes. In this paper we summarise some of the considerations made to address the most crucial issues. For MS who conducted a census only a sample will be released as it is deemed to risky to release the whole population. Also MS who deem too risky the release of FSS microdata might consider sub-sampling as an option for further protection of the original sample. As the geographical location of the agricultural holding is so crucial it seems a preferable strategy the one that maintains, as far as possible, a NUTS2 level and, at the same time, slightly perturbs all the farms rather than concentrating on the protection of the highly visible farms by loosing regional details. By introducing some small flexibilities in the geographical level released, in some areas possible local aggregations could be possible in case of a very reduced number of farms whereas further investigations in some regions could lead to release of data at a lower geographical details.

The choice between releasing the SGM as it is in the original data or recalculate it after the disclosure limitation techniques have been applied (which undoubtedly requires a big effort) is linked to the type of SDC procedures to be put in place. The perturbation of the continuous variables and the aggregation of highly detailed variables could support the possibility of releasing the SGM as it is recoded in the original microdata. For the perturbation part a variation of Individual ranking method has been proposed that copes with the sparsity of the data. As for the aggregation to preserve data utility constraints have been used to maintain the dissemination of *farming type* at *particular* level. As a conclusion of this analysis, we believe there is scope for the development of a microdata file for research for the European FSS.

References

Hundepool, A., Domingo Ferrer, J., Giessing, S., Franconi, L. Spicer, K. Schoulte Nordholt, E. and De Wolf, P.P. (2009). *Handbook on Statistical Disclosure Control*. Available at the Essnet on SDC web site.

Defays, D. and Anwar M.N. (1998). "Masking Microdata Using Micro-Aggregation", Journal of Official Statistics, 14 (4), 449-461.

Eurostat (2003). Handbook on SGM.