List Frames, Area Frames and Administrative Data, Are They Complementary or In Competition?

Elisabetta Carfagna, Professor of Statistics

University of Bologna, Statistics Department, Italy

Abstract: The extraordinary increase of ability to handle and manipulate large sets of data suggests an extensive use of administrative data for saving money, reducing response burden, producing figures for very detailed domains and estimating transition over time. However, statistical systems based on registers have some disadvantages and specific requirements which are analysed in the paper. Some of these disadvantages can be removed if registers are combined with list or area frame sample surveys through calibration estimators or multiple frames methodology.

1. Introduction

Many different data on agriculture are available in the various countries in the world. Administrative data are common almost everywhere, produced on the basis of various data sources. In some countries, a specific data collection is performed with the purpose of producing agricultural statistics, using complete enumeration or sample surveys based on list or area frames (a set of geographical areas) or both. Rationalization is felt as a strong need by many countries, since they deliver different and sometimes non-comparable data; moreover, maintaining different data acquisition systems is very expensive.
We perform an analysis of risks, advantages, disadvantages and requirements of the use of administrative data for statistical purposes. Then, we propose some methods to combine list frames, area frames and administrative data for producing accurate agricultural statistics.

2. Administrative data

In most countries in the world, administrative data on agriculture are available, based on various acquisition systems. Definitions, coverage and quality of administrative data depend on administrative requirements; thus they change as these requirements change. Their acquisition is often regulated by law; thus they have to be collected, independently of their cost, which is very difficult to calculate since most of the work involved is generally performed by public institutions. Main kinds of administrative data relevant for agricultural statistics are records concerning taxation, social insurance and

subsidies. These data are traditionally used for updating a list created by a census. The result is a sampling frame to carry out sample surveys in the period between two successive censuses (most often 4 to 10 years).

The extraordinary increase of ability to handle and manipulate large sets of data, the capacity of some administrative departments to collect data through the web (that allows a very fast data acquisition in standard form) and budget constrains have suggested to explore the possibility to use administrative data more extensively and even to produce statistics through direct tabulation of administrative data.

3. Administrative data versus sample surveys

A statistical system based on administrative data allows to save money and to reduce response burden. It has also advantages that are typical of complete enumeration, such as producing figures for very detailed domains (not only geographical) and estimating transition over time. In fact, statistical units in a panel sample tend to abandon a survey after a while and comparison over time becomes difficult; whilst units are interested or obliged to deliver administrative data.
Various countries in the world are moving from a sample based statistical system to a register based one, where a register is a complete list of objects belonging to a defined objects set and with identification variables that allow to update the register itself.

When a sample survey is performed, first of all, the population is identified, then a decision is taken about: the parameters to be estimated (for the variables of interest and for specific domains) and the levels of accuracy to be reached, taking into account budget constrains.
When statistics are produced on the basis of registers, the procedure is completely different, since data have already been collected. Sometimes objects in the registers are partly the statistical units of the population for which statistics have to be produced and partly something else; thus evaluating under-coverage of registers is difficult.

4. Direct tabulation of administrative data

Two interesting studies (Selander *et al*., 1998 and Wallgren and Wallgren, 1999) were financed jointly by Statistics Sweden and Eurostat. They explored the possibility of producing statistics on crops and livestock through the integrated administrative and control system (IACS, created for European agricultural subsidies) and other administrative data. After a comparison of IACS data with an updated list of farms, the first study came to the following conclusion: “The IACS register is generally not directly designed for statistical needs. The population that applies for subsidies does not correspond to the population of farms which should be investigated. Some farms are outside IACS and some farms are inside this system but do not provide complete information for statistical needs, Supplementary sample surveys must be performed, or the statistical system will produce biased results. To be able to use IACS data for

statistical production the base should be a Farm Register with links to the IACS register.”

4.1. Disadvantages of direct use of administrative data

When administrative data are used for statistical purposes, the first problem to face is that information acquired is not exactly the one needed, since questionnaires are designed for specific administrative purposes. Statistical and administrative purposes require different kinds of data to be collected and different acquisition methods (which strongly influence the quality of data). Strict interaction between statisticians and administrative departments is essential, although it’s not a guaranty that the problem will be solved.
For example, if all the detailed information needed for producing statistics on main crops is acquired through IACS questionnaires, they become very long and complicated and the risk of collecting bad quality data becomes high. Moreover, the acquisition date of IACS data does not allow to collect information concerning yields; thus, an alternative data source is needed for estimating these important parameters.

Administrative data are not collected for pure statistical purposes, with the guaranty of confidentiality and avoiding to use data for other purposes, unless they are aggregated. Administrative data are collected for specific purposes which are very relevant for the respondent such as subsidies or taxation and so on. On one side, this relevance should guaranty accurate answers and high quality of data; on the other, specific interests of respondents can generate biased answers.

For example, IACS declarations have a clear aim; thus applicants devote much attention to records concerning crops with subsides based on surface, due to the controls that are performed, and less attention to surfaces of other crops.
Non clear dynamics can be generated by these controls, since some farmers may decide not to apply for crops with subsides, others may tend to underestimate the surfaces to avoid risks and consequences of controls and others may inflate their declarations, hoping not to be submitted to control.

4.2. Quality control in sample surveys and registers

A pillar of sampling theory is that, when a sample survey is carried out, much care can be devoted to the collection procedure and to data quality control, since a limited amount of data is collected; thus, non sampling errors can be limited. At the same time, sampling errors can be reduced adopting efficient sample designs. The result is that, often, very accurate estimates can be produced with limited amount of data.

The register approach is the opposite: a huge amount of data is collected for other purposes and sometimes a sample of those data is controlled to apply sanctions and

not for evaluating data quality or understanding what can be misleading in the questionnaire and so on.

4.3. Coverage problems of registers

Mentioned studies made an analysis of record linkage results using IACS records and a list of farms created by the census and updated. The telephone number allowed to identify 64.1% of objects in the list of farms and 72.0% of objects in IACS; much better results were achieved associating also other identification variables, such as organisation numbers (BIN) and personal identification numbers (PIN): 85.4% of objects in the list of farms and 95.5% of objects in IACS. However, only 86.6% of objects in IACS and 79% of objects in the list of farms have a one to one match, others have a one to many or many to many match and 4.5% of IACS objects and 14.6 of objects in the list of farms have non match at all.

A comparison of IACS data with estimates derived form a survey of farms has shown that incompleteness of data delivered by some applicants inflates the risk of bias for some crops. In Sweden, they have estimated that for crops with subsidies based on surface and for other crops which are generally cultivated by the same farms, the bias is low, but for other crops it can be about 20%.
Moreover, comparability over time is strongly influenced by the change of the level of coverage in the different years and can give misleading results.

5. Errors in administrative data

Estimation of parameters has a meaning only if the reference population is well defined; while, in most cases, registers are constituted by a series of elements which cannot be considered as belonging to the same population from a statistical viewpoint. For instance, applicant for IACS are not necessarily holders. Therefore, producing statistics about the population of farms requires a very good record linkage process for evaluating coverage problems (see W. Winkler 1995 for a detailed analysis of record matching methods and problems).

Direct tabulation from a register is suggested for a specific variable if the sum of values for that variable presented by all the objects in the register is an unbiased estimator of the total for this variable. This estimator is applied to data affected by errors, since some objects can present inflated values, some others can have the opposite problem; then, some objects that are in the register should not be included and others which are not included should be in the register.

For example, let’s consider IACS declarations for a crop *c*; these data are affected by commission errors (some parcels declared as covered by crop *c* are covered by another crop or their surface is inflated) and omission errors (some parcels covered by crop *c* are not included in IACS declarations or their surface is less than the true). If

commission and omission errors compensate, the sum of declaration for crop *c* is an unbiased estimator of the surface of this crop.

5.1. Quality control of IACS data

An evaluation of commission errors can be made through a quality control on a probabilistic sample of the declarations. Quality control of IACS data is performed every year on the ground on a sample of declarations.
In 2003, at Italian level, for ten controlled crops (or groups of crops) the error was 48,591 ha, 3.9% of controlled surface (1,270,639 ha). For an important crop like durum wheat (national declared surface 1,841,230 ha), 23,314 controls were performed, corresponding to a controlled surface of 347,475 ha (19% of declared surface) and the error was 12,223 ha, 3.5% of controlled surface.
The situation is very different for other crops, such as leguminous, for which the error is 1,052 ha, 16% of controlled surface (6,568 ha). Moreover, if we consider specific geographic domains, for example the area of six provinces out of nine in Sicily, the error for durum wheat in 2000 was 16% of controlled surface.

We cannot say that, at a national level, commission errors for durum wheat amount to 3.9% and reduce the total surface of this percentage for eliminating an upwards bias, because sample selection of IACS quality control is purposive, since its aim is detecting irregularities and not estimating the level of commission errors, thus it tends to be an overestimate of commission errors.
It's evident that quality control is performed for different purposes for statistical surveys and administrative registers and thus gives different results that should not be confused.

6.Commission and omission errors

A possibility to estimate commission and omission errors is given by the study carried out by Consorzio ITA (AGRIT 2000) in Italian Puglia and Sicily regions for durum wheat in 2000. In both regions, an area frame sample survey based on segments with permanent physical boundaries was executed. ITA estimates of durum wheat surfaces were 435,487.3 ha in Puglia, with a coefficient of variation (CV) of 4.8% and 374,658.6 ha in Sicily (CV 5.9%).
Then, for each segment, the surface of durum wheat deriving from the declarations was computed and resulting estimates (IACS estimates) were compared with ITA estimates. IACS estimates where smaller than ITA ones (6.9% less in Puglia and 16.0% in Sicily). Also the sum of IACS declarations (IACS data) was smaller than ITA estimates (10.4% less in Puglia and 12.2% in Sicily).

Parcels declared covered by durum wheat were identified on each sample segment. For some of these parcels, declared surfaces equalled surfaces detected on the ground by the area sample survey; for others, there was a more or less relevant difference. Finally,

these differences were expanded to the universe and estimates of commission errors were produced: 7.8% of the sum of declarations in Puglia and 8.4% in Sicily.
A comparison with ITA estimates suggests the presence of a relevant omission error, that is about 13.9% of ITA estimate in Puglia and 23.3% in Sicily). So high levels of omission error are probably due partly to incorrect declarations and partly to farmers who did not apply.

When data collection is performed for purposes different from pure statistical knowledge and a quality control devoted to identification of irregularities is carried out, declarations can be influenced by complex dynamics, which are difficult to foresee and can produce a bias. Consider that durum wheat is one of the crops with subsidy based on surface, thus considered reliable by mentioned Swedish studies.
Described study also suggests that data for small domains produced by registers can be unreliable due to different dynamics in different domains.

7. Alternatives to direct tabulation

An approach for reducing the risk of bias due to under-coverage of registers and, at the same time, avoiding double data acquisition is sampling farms from a complete and updated list and performing record linkage with the register for capturing register data corresponding to farms selected from the list. If the register is considered unreliable for some variables, related data have to be collected through interviews as well as data not found in the register due to record linkage difficulties

7.1. Matching different registers

Countries with a highly developed system of registers can capture data from the different registers to make comparisons, to validate some data with some others and to integrate them. Of course, very good identification variables and a very sophisticated record linkage system are needed.
Main registers used are the annual income verifications in which all employers give information on wages paid to all persons employed, the register of standardised accounts (based on annual statements from all firms), the VAT register (based on VAT declarations from all firms), the vehicle register (vehicles owned by firms and persons). The combined use of these registers improves the coverage of the population and data quality trough comparison of data in the different registers and allows to describe the socio-economic situation of rural households. However, it doesn’t solve all problems connected with under-coverage and incorrect declaration.

The statistical methodological work to be done for using multiple administrative sources is very heavy (see Wallgren and Wallgren, 1999): editing of data, handling of missing objects and missing values, linking and matching, creating derived objects and variables.

Then, the work to be done for quality assurance is: contacts with suppliers of data, checking of received data, causes and extant of missing objects and values, imputation, causes and extent of mismatch, evaluating objects and variables and reporting inconsistencies between registers, reporting deficiencies in metadata, carrying out register maintenance surveys.
All this is a considerable amount of work, since it has to be performed on the whole registers and its cost is not calculated; moreover, the effect of mismatch or imperfect match or statistical match on statistical estimates is not evaluated.

8. Calibration estimators

A completely different way of taking advantage of registers is the following: the statistical system is based on a probabilistic sample survey with data collected for statistical purposes whose efficiency is improved by the use of register data as auxiliary variable in calibration estimators Deville and Särndal (1992).
Improved efficiency allows to reach the same precision reducing sample size, survey costs and response burden

Consorzio ITA (AGRIT 2000) used IACS data as auxiliary variable in a regression estimator (a kind of calibration estimator). Coefficient of variation (CV) of estimates was reduced from 4.8% to 1.3% in Puglia and from 5.9% to 3.0% in Sicily.
Consider that Landsat TM remote sensed data used as auxiliary variable allowed a reduction of CVs in Puglia to 2.7% and in Sicily to 5.6% (for cost efficiency of remote sensing data see Carfagna 2001b).

When available registers are highly correlated with the variables for which parameters have to be estimated, described approach has many advantages:

1. register data are included in the estimation procedure thus different data are conciliated in one datum;
2. allows a strong reduction of the sample size and thus of survey costs and of respondent burden;
3. if the sampling frame is complete and without duplications there is no risk of under-coverage;
4. data are collected for pure statistical purposes; thus are not influenced and corrupted by administrative purposes.

Disadvantages are that costs and respondent burden are higher than when direct tabulation is performed. A detailed comparison should be made with the costs of a procedure using multiple administrative sources and sample surveys for maintaining registers.
Another disadvantage is the difficulty to produce reliable estimates for small domains, since this approach assumes a small sample size; thus, just few sample units are allocated in small domains and corresponding estimates tend to be unreliable; small area estimation methods should be applied.

9. Combined use of different frames

When various incomplete registers are available and information included in their records cannot be directly used for statistics, a sample survey has to be designed. Most often, administrative data are used for creating one single sampling frame, although on the basis of two or more lists. This approach should be undertaken only if the different lists contribute with essential information to complete the frame and the record matching gives extremely reliable results; otherwise, the frame will be still incomplete and with many duplications.

An alternative approach is treating these registers as multiple incomplete lists from which separate samples can be selected for sample surveys. Then, a two-stage estimator can be adopted, that is an estimator that combines estimates calculated on non-overlapping sample units belonging to the different frames with estimates calculated on overlapping sample units.
This ways of treating different lists does not require record matching of listing units of the different lists. Some two-stage estimators need the identification of identical units only in the overlap samples and some others have been developed for cases in which these units cannot be identified (see Hartley 1962, 1974 and Fuller and Burmeister 1972). Completeness assumption has to be made: every unit in the population of interest should belong to at least one of the frames

9.1 Estimation of a total

For simplicity, let us consider the case of two frames (*A* and *B*), both incomplete and with some duplications, which together cover the whole population. The frames *A* and *B* generate three (2^2-1) mutually exclusive domains: *a* (units in *A* alone), *b* (units in *B* alone), *ab* (units in both *A* and *B*). N_A and N_B are the frames sizes, N_a, N_b and N_{ab} are the domains sizes.
The three domains cannot be sampled directly since we don't know which units belong to each domain and samples of sizes n_A and n_B have to be selected from frames *A* and *B*. Thus n_a, n_{ab}^A, n_{ab}^B and n_b (the subsamples of n_A and n_B respectively which fall into the domains *a*, *ab* and *b*) are random numbers and a post-stratified estimator has to be adopted for the population total.

For simple random sampling in both frames, in case all the domain sizes are known, a post-stratified estimator of the population total is the following:

$$\hat{Y} = N_a \bar{y}_a + N_{ab}(p\bar{y}_{ab}^A + q\bar{y}_{ab}^B) + N_b \bar{y}_b, \qquad (1)$$

where *p* and *q* are non-negative numbers with $p+q=1$; $\bar{y}_a$ and $\bar{y}_b$ denote the respective sample means of domains *a* and *b*; finally, $\bar{y}_{ab}^A$ and $\bar{y}_{ab}^B$ are the sample means of domain *ab*, relative, respectively, to subsamples n_{ab}^A and n_{ab}^B.

$N_a \bar{y}_a$ is an estimate of the incompleteness of frame *B*.

9.2 Accuracy of estimates

Hartley (1962) proposed to use the variance for proportional allocation in stratified sampling as approximation of the variance of the post-stratified estimator of the population total $\hat{Y}$ with simple random sampling in the two frames (ignoring finite population corrections):

$$Var(\hat{Y}) \approx \frac{N_A^2}{n_A}\left[\sigma_a^2(1-\alpha) + p^2\sigma_{ab}^2\alpha\right] + \frac{N_B^2}{n_B}\left[\sigma_b^2(1-\beta) + q^2\sigma_{ab}^2\beta\right], \quad (2)$$

where σ_a^2 σ_b^2 and σ_{ab}^2 are the population variances within the three domains, moreover $\alpha = N_{ab}/N_A$ and $\beta = N_{ab}/N_B$.
Under a linear cost function, the values for $n_A / N_A p$, and n_B / N_B minimising the estimator variance can be determined (see Hartley, 1962).

The knowledge of the domain sizes is a very restrictive assumption that is seldom verified. Often, domain sizes are only approximately known, due to the use of out of date information and lists, that makes difficult to determine whether a unit belongs to any other frame. In such a case, the estimator of the population total given in equation (1) is biased and the bias remains constant as the sample size increases. Many estimators that do not need domain sizes were proposed by Hartley (1962), Lund (1968) and Fuller and Burmeister (1972).

10. Complex sample designs

Generally, complex designs are adopted in the different frames to improve efficiency and this affects the estimators. Hartley (1974) and Fuller and Burmeister (1972) considered the case in which at least one of the samples is selected by a complex design, such as stratified or multistage sampling.

Skinner and Rao (1996) proposed alternative estimators under complex designs where the same weights are used for all the variables. Particularly, they modified the estimator suggested by Fuller and Burmeister for simple random sampling in the two frames, in order to achieve design consistency under complex designs, while retaining the property of being a linear combination of observations and having a simple form.

From a general viewpoint, whatever the sample design in the two frames, using the Horvitz-Thompson estimators of the totals of the different domains, the estimator of the population total is given by:

$$\hat{Y} = \hat{Y}_a + p\hat{Y}_{ab}^A + q\hat{Y}_{ab}^B + \hat{Y}_b. \quad (3)$$

When sample selection is independent in the two frames, the following covariances are zero:

$$Cov(\hat{Y}_a, \hat{Y}_b), Cov(\hat{Y}_a, \hat{Y}_{ab}^B); Cov(\hat{Y}_b, \hat{Y}_{ab}^A); Cov(\hat{Y}_{ab}^A, \hat{Y}_{ab}^B), \quad (4)$$

and the variance of population total in equation (3) is:

$$Var(\hat{Y}) = Var(\hat{Y}_a) + p^2 Var(\hat{Y}_{ab}^A) + (1-p)^2 Var(\hat{Y}_{ab}^B) + Var(\hat{Y}_b) + \\ + 2pCov(\hat{Y}_a, \hat{Y}_{ab}^A) + 2(1-p)Cov(\hat{Y}_b, \hat{Y}_{ab}^B). \quad (5)$$

Thus, the value of *p* that minimises the variance in equation (5) is:

$$p_{opt} = \frac{Var(\hat{Y}_{ab}^B) + Cov(\hat{Y}_b, \hat{Y}_{ab}^B) - Cov(\hat{Y}_a, \hat{Y}_{ab}^A)}{Var(\hat{Y}_{ab}^A) + Var(\hat{Y}_{ab}^B)}. \quad (6)$$

The optimum value for *p* is directly related to the precision of $\hat{Y}_{ab}^A$.

When frame *A* is complete, $\hat{Y}_b$ in equation (3) is zero as well as $Cov(\hat{Y}_b, \hat{Y}_{ab}^B)$ in equations (5) and (6); thus, we have:

$$\hat{Y} = \hat{Y}_a + p\hat{Y}_{ab}^A + q\hat{Y}_{ab}^B \quad (7)$$

with variance:

$$Var(\hat{Y}) = Var(\hat{Y}_a) + p^2 Var(\hat{Y}_{ab}^A) + (1-p)^2 Var(\hat{Y}_{ab}^B) + 2pCov(\hat{Y}_a, \hat{Y}_{ab}^A) \quad (8)$$

11. Area frames

When completeness is not guaranteed by the combined use of different registers, an area frame should be adopted for avoiding bias, since an area frame is always complete, and remains useful a long time (Carfagna 1998).
The completeness of area frames suggests their use in many cases:

1. if other complete frame is not available;
2. if an existing list of sampling units changes very rapidly;
3. if an existing frame is out of date;
4. if an existing frame was obtained from a census with low coverage;
5. if a multiple purpose frame is needed for estimating many different variables (agricultural, environmental etc.).

Area frame sample designs also allow objective estimates of characteristics that can be observed on the ground, without interviews. Besides, the materials used for the survey and the information collected help to reduce non sampling errors in interviews and are a good basis for data imputation for non-respondents; finally, the area sample survey materials are becoming cheaper and more accurate.

Area frame sample designs also have some disadvantages, such as the cost of implementing the survey program, the necessity of many cartographic materials, the sensitivity to outliers and the instability of estimates. If the survey is conducted through interviews and respondents live far from the selected area unit, their identification may be difficult and expensive, and missing data tend to be relevant.

12. Combining a list and an area frame

The most widespread way to avoid instability of estimates and to improve their precision is adopting a multiple frame sample survey design. For agricultural surveys, a list of very large operators and of operators that produce rare items is combined with the area frame. If this list is short, it is generally easy to construct and update. A crucial aspect of this approach is the identification of the area sample units included in the list frame. When units in the area frame and in the list sample are not detected, the estimators of the population totals have an upwards bias.

Sometimes, a large and reliable list is available. In such cases, the final estimates are essentially based on the list sample. The role of the area frame component in the multiple frame approach is essentially solving the problems connected with incompleteness of the list and estimating the incompleteness of the list itself.
In these cases, updating the list and record matching for detecting overlapping sample units in the two frames are difficult and expensive operations that can produce relevant nonsampling errors (Vogel 1975 and Kott and Vogel 1995).

Combining a list and an area frame is a special case of multiple frame sample surveys in which sample units belonging to the lists and not to the area frame do not exist (domain *b* is empty) and the size of domain *ab* equals N_B (frame *B* size: the list size, that is known).
This approach is very convenient when the list contains units with large (thus probably more variable) values of some variable of interest and the survey cost of units in the list is much lower than in the area frame (Kott and Vogel 1995; Carfagna, 2001b).

The optimum value of *p* in equation (6) depends on the item and can assume very different values for the different variables. In most applications, the value of *p* is chosen equal to zero and the resulting estimator is called screening estimator, since it requires the screening and elimination from the area frame of all the area sampling units included in the list:

$$\hat{Y} = \hat{Y}_a + \hat{Y}_{ab}^B , \qquad (9)$$

and its variance is:

$$Var(\hat{Y}) = Var(\hat{Y}_a) + Var(\hat{Y}_{ab}^B) \qquad (10)$$

13. Conclusions

The extraordinary increase of ability to handle and manipulate large sets of data suggests to explore the possibility to use administrative data more extensively and even of creating statistical systems based on administrative data. This is a way for saving money, reducing response burden producing figures for very detailed domains and allowing estimation of transition over time.

However, definitions, coverage and quality of administrative data depend on administrative requirements; thus they change as these requirements change. Then information acquired is not exactly the one needed for statistical purposes and sometimes objects in the registers are partly the statistical units of the population for which statistics have to be produced and partly something else; thus evaluating under-coverage is difficult.

Administrative departments collect data for specific purposes and perform quality controls for detecting irregularities and not for evaluating the data quality. These quality controls can be misleading if used for estimating commission errors and declarations can be influenced by complex dynamics, which are difficult to foresee and can produce a bias.

Comparability over time is strongly influenced by the change of the level of coverage in the different years.
The combined use of registers improves the coverage of the population and data quality through comparison of data in the different registers and allows to describe socio-economic situations of rural households.
Very good identification variables and a very sophisticated record linkage system are needed and a heavy statistical methodological work has to be done. The effect of imperfect matching on the estimates of parameters should be evaluated.

An approach for reducing the risk of bias due to under-coverage of registers and, at the meant time, avoiding double data acquisition is sampling farms from a complete and updated list and performing record linkage with the register for capturing register data corresponding to farms selected from the list. If the register is considered unreliable for some variables, related data have to be collected through interviews as well as data not found in the register due to record linkage difficulties

A completely different way of taking advantage of registers is improving the efficiency of estimates based on a probabilistic sample survey by the use of register data as auxiliary variable in calibration estimators Deville and Särndal (1992).
Improved efficiency allows to reach the same precision reducing sample size, survey costs and response burden

When various incomplete registers are available but information included in their records cannot be directly used and a sample survey has to be designed, these registers can be treated as multiple incomplete lists from which separate samples can be selected. This way of treating different lists does not require record matching of listing units of the different lists.

When completeness is not guaranteed by the different registers, an area frame should be adopted for avoiding bias, since an area frame is always complete, and remains useful a long time.
The most widespread way to avoid instability of estimates based on an area frame and to improve their precision is adopting a multiple frame sample survey design that

combines the area frame with a list of very large operators and of operators that produce rare items. This approach is very convenient when the list contains units with large (thus probably more variable) values of some variable of interest and the survey cost of units in the list is much lower than in the area frame

The author thanks very much many people in the following organisation for their support: Eurostat, Italian Ministry of Agriculture and Forest, Istat, Consorzio ITA, AGEA, Statistics Sweden, UNECE and Statistics Department of the University of Bologna.

References

Amrhein J., Hicks S., Kott P. (1996), Methods to Control Selection when Sampling from Multiple List Frames, *Proceeding of the Section on Survey Research Methods*, American Statistical Association.
Armstrong B. (1979), Test of multiple frame sampling techniques for agricultural surveys: New Brunswick, 1978, *Proceeding of the Section on Survey Research Methods*, American Statistical Association, pp. 295-300.
Baldwin, J. , Dupuy, R. , and Penner, W. (1992), ``Development of longitudinal panel data from business registers: Canadian experience (STMA V34 4759)'', *Statistical Journal of the U.N. Economic Commission for Europe*, **9** , 289-303
Blom, E. , and Carlsson, F. (1999), ``Integration of administrative registers in a statistical system: A Swedish perspective'', *Statistical Journal of the U.N. Economic Commission for Europe*, **16** , 181-196
Bailey J.T., Kott P. S. (1997), An application of Multiple List Frame Sampling from Multi-purpose Surveys, *Proceeding of the Section on Survey Research Methods*, American Statistical Association, pp. 496-500.
Balicki, A. , and Szreder, M. (1997), ``Usefulness of official registers in sample surveys in Poland (STMA V39 4628)'', *Statistics in Transition*, **3** , 315-328
Bankier M. D. (1986), Estimators Based on Several Stratified Samples With Applications to Multiple Frame Surveys, *Journal of the American Statistical Association*, 81, 1074-1079.
Biffignandi, Silvia , and Butti, Christine (1993), ``Administrative registers and national surveys'', *Proceedings of the International Conference on Establishment Surveys. Survey Methods for Businesses, Farms, and Institutions*, 542-547
Bosecker R. R. and Ford B. L. (1976), Multiple Frame Estimation with Stratified Overlap Domains, *Proceedings of the Social Statistics Section*, American Statistical Association, Part I, pp. 219-224.
Brackstone G. J. (1987) “Issues in the use of administrative records for statistical purposes”, Survey Methodology, 13, pp. 29.43.
Brackstone G. J. (1999) “Managing data quality in a statistical agency”, Survey Methodology, 25, pp. 139-149.
Carfagna, E. (1998). Area frame sample designs: a comparison with the MARS project, Proceedings of Agricultural Statistics 2000, International Statistical Institute, Voorburg. pp. 261-277.
Carfagna E. (2001a), “Multiple Frame Sample Surveys: Advantages, Disadvantages and Requirements'', in International Statistical Institute, Proceedings, Invited papers, International Association of Survey Statisticians (IASS) Topics, Seoul August22-29, 2001.
Carfagna, E. (2001b). Cost-effectiveness of remote sensing in agricultural and environmental statistics, *Proceedings of the Conference on Agricultural and Environmental Statistical Applications in Rome (CAESAR).* June 5-7, Vol. 3 pp. 618-627. http://www.ec-gis.org/
Clark, Cynthia Z. F. , and Vacca, Elizabeth Ann (1993), ``Ensuring quality in U.S. agricultural list frames'', Proceedings of the International Conference on Establishment Surveys. Survey Methods for Businesses, Farms, and Institutions, 352-361
Colledge M. J. (1995), Frames and Business Registers: An Overview in Cox, Binder, Chinnapa, Christianson, Colledge, Kott (Eds), *Business survey methods*, Wiley, New York, pp. 21-47.
Consorzio ITA (2000), AGRIT 2000 Innovazione tecnologica. Studio per l’integrazione dati ADRIT-PAC, Ministero delle Politiche Agricole e Forestali.

Eurostat (1997), Proceedings of the seminar on the use of administrative sources for statistical purposes : Luxembourg, 15-16 January 1997,Luxembourg : Office for official publications of the European Communities.
Fienberg S.E., Johnson M.S., Junker B.W. (1999), Classical Multilevel and Bayesian Approaches to Population Size Estimation Using Multiple Lists, *Journal of the Royal Statistical Society*, vol. 162, Part 3, pp. 383-405.
Fuller W. A., Burmeister L. F. (1972) Estimators for samples from two overlapping frames, *Proceedings of the Social Statistics Section*, *American Statistical Association*, pp. 245-249.
Goldberg M. L., Gargiullo P. M. (1988), Variance Estimation Using Pseudostrata for a List-supplemented Area Probability Sample, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 479-484.
Groves R. M., Lepkowski J. M. (1985), Dual Frame, Mixed Mode Survey Design, *Journal of Official Statistics*, 1, pp. 263-286.
Haines, Dawn E. , Pollock, Kenneth H. , and Pantula, Sastry G. (2000), ``Population size and total estimation when sampling from incomplete list frames with heterogeneous inclusion probabilities'', Survey Methodology, 26 (2) , 121-129
HANSEN M., HURWITZ W., MADOW W. (1953) *Sample Survey Methods and Theory*, New York, John Wiley and Sons, vol. I (pp. 515-558).
Hartley H. O. (1962), Multiple-frame surveys, *Proceedings of the Social Statistics Section*, *American Statistical Association*, pp. 203-206.
Hartley H. O. (1974), Multiple Frame Methodology and Selected Applications, *Sankhya*, vol.36, series C, Pt.3, pp. 99-118.
Kalton G., Anderson D. W. (1986), Sampling rare populations, *Journal of the Royal Statistical Society*, Ser. A, 149, pp. 65-82.
Jansson, K. (1994), ``Use of administrative registers for income statistics in Sweden'', *Statistical Journal of the U.N. Economic Commission for Europe*, **11** , 211-228
Jensen, P. (1994), ``Business registers for statistical use: The case of Denmark'', *Proceedings of the Italian Statistical Society, Volume 1*, **1** 303-314
Kott, Phillip S. , Amrhein, John F. , and Hicks, Susan D. (1998), ``Sampling and estimation from multiple list frames'', Survey Methodology, 24 , 3-9
Kott P. S., Bailey J. T. (2000), The Theory and Practice of Maximal Brewer Selection With Poison PRN Sampling, *International Conference on Establishment Surveys – II, Survey Methods for Business, Farms and Institutions*, American Statistical Association, June 17-21, 2000.
Kott P. S., Vogel F. A. (1995), Multiple-frame business surveys, in Cox, Binder, Chinnapa, Christianson, Colledge, Kott (Eds.), *Business survey methods*, Wiley, New York, pp. 185-201.
Lehtonen, Risto , and Veijanen, Ari (1999), ``Use of register data to improve the estimation in a sample survey: The Finnish Labour Force Survey as a case study'', *Statistics, Registries, and Science. Experiences from Finland*, 197-210
Lepkowski J., Groves R. M. (1986), A Mean Squared Error Model for Dual Frame Mixed Model Survey Design, *Journal of the American Statistical Association*, 81, 930-937.
Lohr S. L., Rao J. N. K. (2000), Inference From Dual Frame Surveys, *Journal of the American Statistical Association*, vol. 95, No. 449, Theory an Methods, pp. 271-280.
Lund R. E. (1968) Estimators in Multiple Frame Surveys, *in Proceedings of the Social Statistics Section*, American Statistical Association, pp. 282-288.
Longva, Svein , Thomsen, Ib , and Severeide, Paul Inge (1998), ``Reducing costs of censuses in Norway through use of administrative registers'', *International Statistical Review*, **66** , 223-234
Mamberti Pedullà, M. G. (1994), ``Fiscal registers and national accounts (Italian)'', *Proceedings of the Italian Statistical Society, Volume 1*, **1** 327-338
Martini, Marco (1993), ``Statistical aspects of business registers integration'', *Proceedings of the International Conference on Establishment Surveys. Survey Methods for Businesses, Farms, and Institutions*, 536-541
Myrskylä, P. (1999), ``New statistics made possible by the use of registers'', *Statistical Journal of the U.N. Economic Commission for Europe*, **16** , 165-180
Ohlsson E. (1995), Coordination of Samples Using Permanent Random Numbers, in Cox, Binder, Chinnapa, Christianson, Colledge, Kott (Eds.), *Business survey methods*, Wiley, New York, pp. 153-169.

Rao J. N. K., Wu C. F. J. (1985), Inference From Stratified Samples: Second-Order Analysis of Three Methods for Nonlinear Statistics, *Journal of the American Statistical Association*, vol. 80, n. 391, pp. 620-630.
Selander R., Svensson J., Wallgren A., Wallgren B. (1998), *How should we use IACS data?*, Statistics Sweden.
Spears, Floyd M. , Chhikara, Raj S. , and Perry, Charles R. (1998), ``An investigation of incompleteness of list frames in US agricultural surveys'', ASA Proceedings of the Section on Survey Research Methods, 505-510
Skinner C. J. (1991), On the Efficiency of Raking Ratio Estimation for Multiple Frame Surveys, *Journal of the American Statistical Association*, vol. 86, No. 415, Theory an Methods, pp. 779-784.
Skinner C. J., Holmes D. J., Holt D. (1994) Multiple Frame Sampling for Multivariate Stratification, *International Statistical Review*, 62, 3, pp.333-347.
Skinner C. J., Rao J. N. K. (1996), Estimation in Dual Frame Surveys With Complex Designs, *Journal of the American Statistical Association*, 91, 349-356.
Thompson, S. K., Seber G. A. F. (1996), *Adaptive sampling*, Wiley, New York.
Tuinen, H. K. , Van, Altena , and Imbens, H. C. M. (1994), ``Surveys, registers, and integration in social statistics'', *Statistical Journal of the U.N. Economic Commission for Europe*, **11** , 321-345
Vogel F. A. (1975), Surveys with Overlapping Frames - Problems in Applications*, Proceeding of the Social Statistics Section, American Statistical Association*, pp. 694-699.
Wallgren A., Wallgren B. (1999), *How can we use multiple administrative sources?*, Statistics Sweden
Winkler W. E. (1995), Matching and Record Linkage, in Cox, Binder, Chinnapa, Christianson, Colledge, Kott (Eds), *Business survey methods*, Wiley, New York, pp. 355-384.
Yung W., Rao J. N. K. (1996), Jackknife Linearization Variance Estimators Under Stratified Multi-Stage Sampling, *Statistics Canada*, Vol. 22, n. 1, pp. 23-31.