Coverage issues in agriculture statistics when using administrative data

Tiziana Tuoto
Italian Institute of Statistics, Department for Methodologies
Via Balbo, 16
Rome, Italy
tuoto@istat.it

DOI: 10.1481/icasVII.2016.f32e

ABSTRACT

Resumé: Administrative sources provide huge amount data that could be directly or indirectly introduced in several phases of the statistical process. The use of administrative data for statistical purpose often requires matching records, due to the fact that an administrative source alone is unlikely to be enough to produce high quality results with negligible or admissible errors. One of the first purposes of combining administrative data is to collect a statistical register of the target population or at least a master frame for sampling: in some developed countries, census has been totally substituted by a set of high quality administrative registers. In general, the use of administrative data in official statistics seems a crucial step for modernizing the existing processes. This paper focalises the coverage aspect connected to the use of administrative data in census context, in particular because in this case the perspective is moving from the traditional census under-count issues to the strong relevance of over-count related to the use of administrative sources.

Keywords: Coverage errors, Register and sampling frame, Linkage errors

1. Introduction

Nowadays administrative data represent a great opportunity for agricultural statistics: their use allows to meet the increasing needs of knowledge and information even at local level, by means of already available data, without further collection cost and response burden. Administrative sources provide huge amount data that could be directly or indirectly introduced in several phases of the statistical process in order to improve the overall quality of information. However, the complete and conscious utilization of administrative data for statistical purpose requires to positively answer to questions related to their quality according to the statistical perspective and the effect of their usage on the overall quality of the statistical results.

The use of administrative data for statistical purpose often requires matching records, due to the fact that an administrative source alone is unlikely to be enough to produce high quality results with negligible or admissible errors. One of the first purposes of combining administrative data is to collect a statistical register of the target population or at least a master frame for sampling: in some developed countries, census has been totally substituted by a set of high quality administrative registers. In general, the use of administrative data in official statistics seems a crucial step for modernizing the existing processes. In evaluating the impact of the use of administrative data when building a statistical register (both for substituting traditional census or for preparing a master frame for sampling), with respect to the coverage issues, the perspective is moving towards the strong relevance of over-count of administrative sources. In fact, administrative data essentially ensure a broad coverage of the interest population thanks to several separate sources, even if a little degree of under-coverage for particular population subgroups could still remain. Conversely, over-coverage is expected to be the more substantial problem. Further coverage problems could be generated by the record linkage procedure utilized in order to match all the information related to the same units reported in the different data sources. This could dramatically happen when administrative sources refer to units that are associated only indirectly to the statistical units of interest.

This paper focuses on the effect of administrative source use, with particular attention devoted to over-coverage errors and matching errors. The work reports examples from the 2010 Italian Agricultural Census. In this context the target is to estimate the unknown total amount of farm

2. The coverage issues

Data collection is inevitably subject to error, whatever is the method used to collect it. Therefore, it is necessary to devise a mechanism to assess the quality and reliability of data sources, in order to make robust adjustments and to improve the quality of the resulting estimates. Coverage error in data collection can be classified in two main groups: under-coverage (units which are completely missed by the data collection process but should be included in the population definition) and over-coverage (additional erroneous units which appear in the collected data source but which should not be included in the population). For agricultural census, example of over-coverage might include, together with duplicate records, big kitchen garden erroneously classified as little farm; farm not more operating (death farm) but not cancelled on the list.

In the traditional census, coverage issue is primarily concerned with estimating and adjusting for under-coverage. This is generally achieved through an independent coverage survey (the PES, Post-Enumeration Survey) and a capture-recapture or Dual System model (Wolter, 1986) aiming at estimating those units which have not been enumerated both by the census and the survey. The capture-recapture (CRC in the following) model is subject to some strong assumptions:

1. the interest population is closed;
1. records from both data sources can be matched without errors;

3. units have the same capture probabilities within each sources (homogeneity probability assumption);
4. over-count in both sources is negligible;
5. the two sources of data are independent.

In formula, say N_{11} the units enumerated in both the census and the PES, N_{1+} the units recorded in the census and N_{+1} those captured in the PES, the Dual System estimator (DSE in the following) of the unknown total amount of the population N is:

$$\hat{N}_{DSE} = \frac{N_{1+} \times N_{+1}}{N_{11}} \quad (1)$$

Several extensions and adjustments of the DSE have been proposed over the time in order to avoid biases due to any failure of these assumptions – resulting in the DSE under-estimates or over-estimates of the true total amount of the population. For instances, without being exhaustive, the main methods used to include dependencies between sources and heterogeneity of captures are extensions of log-linear model (Fienberg 1972; Cormack 1989; Chao 2001, Agresti 1994), conditional multinomial logit model (Chen and Kuo 2001; Zwane and van der Heijden 2005), latent class model (Bartolucci and Forcina 2005) and Bayesian capture-recapture model (Ghosh and Norris 2005).

As far as the matching errors is concerned, Ding and Fienberg (1994) propose simple method to achieve "linkage error unbiased" estimates of population total by explicitly modeling the two types of matching errors. Generally, linkage errors are commonly subdivided in two groups: false matches when records referred to different entities are linked, and missing matches when records that really represent the same unit are erroneously left unmatched. False matches directly deflate the population totals, causing under-coverage of the census list, while missing matches directly inflate the population totals, causing over-coverage of the census list.

Recent works take into consideration bias due to the over-coverage of sources, particularly relevant when dealing with administrative data, affected by significant level of list inflation due to different reasons. Meeting the assumption of close population should help to reduce the over-coverage. It is clear that matching errors will determine bias in over-coverage adjustments, with inflation of over-coverage in case of missing matches and deflation of over-coverage in case of false matches. More and more complicated is the case in the farm field, because of farm demography, changing in organization and operational status, difficulties in verifying farm definition for small farms, etc. All these factors directly and dramatically affect the linkage results, with consequence impact on over-coverage assessment.

Large et al. (2011) propose an adjustment to the DSE in order to correct the census totals by the over-coverage components. For household and population census, they consider four types of over-count:

- duplicate returns at the same location (for example, a paper and internet return for the same household)
- duplicate returns from different locations (for example, students at their term-time and parents' address)
- a return counted in the wrong place (local over-count),
- erroneous returns (units which should not be in the census population such as a baby born just after census day and included on the return).

Therefore, they estimate the correct number of census returns by both local and duplicate over-count and include it in the DSE.

The US Census Bureau carries out a further sample survey (the E sample) in the same areas of the PES in order to estimate erroneous records, that should not have been collected in the census. Due to US level of erroneous enumeration, a dedicated survey is considered in order to determine the correct enumeration, via field work in E sample survey. Therefore, the erroneous records are removed before DSE is calculated. In 2010, the estimation procedure uses logistic regression. These models allow the use of more variables, and therefore produce better estimates of percentage undercount by single demographic characteristics. Three logistic regression models were developed, according to the following dependent variables:

- the probability of being a 'data-defined' enumeration (the data source for estimating the model parameters is all census people in housing units)
- the probability of being a correct enumeration (the data source for estimating the model parameters is the people in the E sample)
- the probability of being matched (the data source for estimating the model parameters is the people in the P sample)

Predictions are then made by inputting the values of the independent variables for each census case into the estimated models. This enables the prediction of correct enumeration rate (with E sample) for census and match rate (with P sample) for census at person level and use these predicted values in the DSE estimation. The approach adopted for the E sample seems the most feasible to view out-of-date or incorrect administrative source records, that could be considered being equivalent to erroneous enumerations in the US census. This means to carry out two separate independent under-coverage and over-coverage surveys, and their records should be matched each other's prior to DSE being calculated.

3. Lesson learned by the 2010 Census of Agriculture in Italy

Coverage issues related to the usage of administrative archives for census purpose have been dramatically experienced in carrying out the PES for the 2010 Italian Agricultural Census. The last Italian agricultural census was register-assisted, that is a preliminary list was provided before the census field operations in order to assist to census mail out. This preliminary list was obtained by matching at unit level several (more then 10) agriculture-related lists: the main ones were agricultural benefit registers, fiscal registers, chambers of commerce registers, cadastral archive, the previous agricultural census list. The goal was to build a list of farms as complete as possible; each record on the list includes name, address, and telephone number of the farm operator plus additional information on the agricultural place, as the address of the farm headquarter and the size in acres of the farm. Most names on a newly acquired list were already on the building frame so those found were set aside and a score of activity status were assigned to each unit. A huge matching work has been done in order to recognize the same farm across the several administrative registers, each of them may report only partial information on the farm size, the farm operator, the land owner or someone involved in the farm operation (for instance a member of the family, a partner, an administrator…); however, with respect to the most ambiguous case, the pre-census list were built admitting some over-coverage. Then, the Census proceeded with field operations on the basis of this preliminary list.

Independently from the census and its preliminary list, to measure the census coverage errors, a PES was designed keeping in mind mainly the under-coverage issue, as usual in the previous census round. However, the linkage step of the PES data highlighted some evidence with respect

over-coverage and further record linkage activities were carried out, matching PES data with all the administrative sources which contributed to the pre-census list. This complex linkage procedure allows to identify the main components of under- and over-coverage. The former ones were:

1. "non-response or kitchen garden", farms included in the census list but hesitated as non-respondent or refusal or not eligible due their size (i.e. kitchen garden) after the census operation;
2. "in pre-census list", farms that were in the pre-census list, but were classified as not eligible before the Census operations;
3. "in administrative archives", farms stemmed in one of the original administrative sources but that have been excluded in the pre-census list;
4. "not present in any lists", PES farms not linked with any units in Census, in pre-census list or in others available archives;

While the forth category represents under-coverage with respect to all the available information, the previous ones explain the under-coverage according to lists as less focused on farms as the number of the category increases.

As far as the over-coverage is concerned, the category "farm with multiple links" was used to represent the over-coverage risk, identifying those farms for which the linkage procedure proposed as correct link more than one unit in the census. Those multiple links (from n units in one file to m units in the other) couldn't be solved according to any optimization algorithm (as in standard record linkage problems) because they represent actual mutual connections related to the land ownership or the farm operation conducted by several different relatives of the same family, on the same land area, at least on the basis of the available information.

Moreover, the category "farm demography" was introduced to measure the risk of linkage errors. In fact, strictly speaking the "farm demography" should not represent over-coverage, because continuity rules are respected in order to assure that the same unit has been identified. However, for these farms the linkage status is less positively defined, further investigations on different variables with respect to those utilized for the linkage were needed, in order to assess the continuity rules hold in despite of the strong changes in the demographic variables (names, organization, land ownership, land extension). So, according to the linkage point of view, the "farm demography" gives a measure of criticisms in identifying farms connected to the use of the administrative sources.

The previous phenomena can be summarized in the following tables, where figures on their impact are reported.

Table 1: *Percentage of under-coverage factors with respect to the total amount of under-coverage*

Cause of under-coverage	Relative Percentage
1. non response or kitchen garden	19.25
2. in pre-census list	5.52
3. in administrative archives	41.65
4. not present in any lists	33.58

Table 2: *Percentage of over-coverage and linkage errors with respect to the total amount of covered farms*

Over-coverage and linkage errors	Relative percentage on the total amount of covered farms
a. farm demography	5.51
b. farm with multiple links	1.91

Table 1 shows that the hardest component of under-coverage (farms not present in any lists) is just one third of the total amount. Most of the not-covered farms (41%) were remain in the administrative sources used to build the pre-census list, meaning that if the procedures adopted in forming the preliminary list is improved enough, the total census under-coverage will be halved. Conversely, the percentage of farms remained in the pre-census list is small (about 5%), while actual farms but differently assigned in the census list is a most considerable percentage (about 20%) with respect to the overall under-coverage. Finally, table 2 shows that the risk of linkage errors, represented by the "farm demography" that explain the difficulties in linking units connected to the same farms from the several administrative sources containing so different kinds of entities, seems bigger than the risk of over-coverage, explained by the "farm with multiple links".

The Table 2 provides measures of over-coverage and linkage errors manly related to the use of administrative data in preparing census list. These errors should be introduced in the selected estimator for the evaluation of the total amount of the interest population in order to reduce its bias. Nevertheless, as far as the 2011 Italian Agricultural PES is concerned, no adjustments were applied and these measurements were provided mainly to stress their importance in a future scenario where the traditional census would be substituted by a farm register built exploiting the administrative sources. In fact, it is important to underline the PES was not designed in order to measure and correct the over-coverage, but only to evaluate the under-coverage of the census list. Anyway, in the next paragraph, an exercise is proposed in order to evaluate the impact of linkage and over-coverage errors on the standard DSE results.

3.1 Adjusting for over-coverage and linkage errors: an exercise

In order to evaluate the effect of over-coverage, the census counts can be deflated by the estimated number of over-enumerated farms, obtained by means of the over-coverage risk represented by "farm with multiple links" reported in table 2. The approach is similar to those applied in UK and Israeli, with simplifications due to the reasonable assumption of absence of local over-coverage (farms are likely less and less moveable than people). In formula,

$$\hat{N}^{O}_{DSE} = \frac{(N_{1+} - \hat{O}) \times \hat{N}_{+1}}{\hat{N}_{11}} \qquad (2)$$

where the DSE estimator of the population total is corrected by the over-coverage error and the number of over-enumerated farms O, the PES totals and the units enumerated both by census and PES are sample estimates.

On the other hand, the adjustment for linkage errors can be obtained following the Ding and Fienberg (1994) proposal. In this context, the probability of missing a true link (i.e. the false non-matches errors, when records that really represent the same unit are erroneously left unmatched) can be considered negligible and approximated by zero (α=1, according to the Ding and Fienberg notation). Otherwise, the probability of false matches (when records referred to different entities are linked) can be evaluated by the risk of linkage errors represented by the "farm demography" (β=0.0551 in the Ding and Fienberg notation). The resulting "linkage error adjusted" DSE total estimator is:

$$\hat{N}^{L}_{DSE} = \frac{(N_{1+} + \hat{N}_{+1} - \hat{N}_{11})}{\hat{p}^{L}_{1+} + \hat{p}^{L}_{+1} - (1-\beta)\hat{p}^{L}_{1+}\hat{p}^{L}_{+1} - \beta\hat{p}^{L}_{1+}} \quad (3)$$

Moreover, the over-coverage and the linkage errors can be taken both into account, deflating the census totals of the estimated over-enumerated units and re-calculating the $\hat{p}^{L}_{1+}$ and $\hat{p}^{L}_{+1}$ as well. In this way, the following linkage and over-coverage unbiased estimator will be obtained:

$$\hat{N}^{L-O}_{DSE} = \frac{(N_{1+} - \hat{O} + \hat{N}_{+1} - \hat{N}_{11})}{\hat{p}^{L-O}_{1+} + \hat{p}^{L-O}_{+1} - (1-\beta)\hat{p}^{L-O}_{1+}\hat{p}^{L-O}_{+1} - \beta\hat{p}^{L-O}_{1+}} \quad (4).$$

In table 3, the resulting under-coverage ratios corresponding to these different estimators are reported, as well as the percentage relative difference with respect to the standard DSE estimate.

Table 3: *The coverage ratio estimates and percentage relative difference with respect to the standard DSE estimate*

	Under-coverage ratio estimates	Percentage Relative Difference
$\hat{\tau}_{DSE}$	11.70	--
$\hat{\tau}^{O}_{DSE}$	9.98	14.73
$\hat{\tau}^{L}_{DSE}$	11.44	2.21
$\hat{\tau}^{L-O}_{DSE}$	9.60	17.93

The table shows clearly that the adjustments for over-coverage and linkage errors work both in the same direction, decreasing the under-coverage rate. Moreover, even if the linkage errors risk seems bigger than the risk of over-coverage (table 2), their impact on the DSE adjustment is exchanged, the over-coverage adjustment being more effective than the linkage one. Finally, it seems important to remark that in this adjusting estimation exercise the measures of over-coverage and linkage errors were derived both from the PES, that was designed uniquely for evaluating the under-coverage. Due to the relevance of over-coverage issues when dealing with administrative sources, it could be preferable to achieve more reliable measures of this kind of error by means of explicit and specific methodologies, as a dedicated survey (like in US) or detailed analyses on a designed sample for quality assessment.

4. Concluding remarks and future works

This paper focalizes attention on quality issues in using administrative data, mainly stressing under-coverage, over-coverage and imperfect matching errors. Examples are reported, from the Post-Enumeration Survey of the 2010 Italian Agricultural Census. In this context the target is to estimate the unknown total amount of farm. The case is complicated by difficulties in recognizing the farm entity in the several administrative sources exploited for assisting the census, with consequent linkage problems in the resulting pre-census list. Other issues are related to changes and modification in the farm entities, the farm demography, also reflecting in linkage errors when comparing data sources. In the framework of capture-recapture models, generally adopted to evaluate the unknown total amount of a population, these facts represent removals of basic assumptions, i.e. the closeness of the population, the absence of over-coverage and the perfect matching hypotheses. Standard estimators should be adjusted in order to take into account the

potential biases introduced when these assumptions do not hold, as already done in case of removal or relaxing the hypotheses of independence of captures and homogeneity of capture probabilities.

The analyses proposed in this paper try to recognize the main components of coverage errors so to identify possible solutions. Generally speaking, when administrative data are exploited for statistical purpose, strategies for minimising over-coverage are adopted, as well as for reducing matching errors. However, even if over-coverage is reduced there will likely be some unknown residual amount which the estimation methodology has to take account for. Moreover, minimising over-coverage risks increasing under-coverage and the robustness of the dual system estimation approach to departures from its assumptions tends to decline as under-coverage increases and this could risk unanticipated errors. It seems possible that the effect of matching errors and over-coverage could be in the same direction, leading to over-estimation of the population, due to the fact that even relatively small errors could lead to non-negligible biases. So further investigation in the estimation methodology is needed in order to adequately take account of these kinds of errors.

REFERENCES

Agresti A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort. Biometrics 50, 494-500

Bartolucci F., Forcina A. (2006). A class of latent marginal models for capture-recapture data with

continuous covariates, Journal of the American Statistical Association, 101, pp. 786-794

Brown J., Taylor A. and Abbott O. (2009). 'Overcount in the 2011 Census: Estimation Issues' proceedings of the 17th meeting of the Government Statistical Service Methodology Advisory Committee, Office for National Statistics

Chao A. (2001). An overview of closed Capture-Recapture Models, Journal of Agricultural, Biological, and Environmental Statistics, 6, pp. 158-175

Chen Z., Kuo L. (2001). A note on the estimation of the multinomial Logit Model with Random effects, The American Statistician, 55, pp. 89-95

Cormack, R. M. (1989). Log-linear models for capture-recapture. Biometrics 45, 395-413

Ding, Y., Fienberg, S.E. (1994). Dual system estimation of Census undercount in the presence of matching error, Survey Methodology, 20, 149-158

Fienberg S. E. (1972). The multiple recapture census for closed populations and incomplete 2k contingency tables. Biometrika 59, 3, p. 591

Ghosh S.K., Norris J.L. (2004). Bayesian Capture-Recapture Analysis and Model Selection Allowing for Heterogeneity and Behavioral Effects, NCSU Institute of Statistics, Mimeo Series 2562, pp. 1-27

Large, A., Brown, J., Abbott, O. and Taylor, A. (2011). Estimating and Correcting for Over-count in the 2011 Census. Survey Methodology Bulletin, 69, 35-48

Wolter, K.M. (1986). Some coverage error models for census data. Journal of the American Statistical Association Vol. 81, No. 394, pp338-346

Zwane, E., Van Der Heijden P. (2005). Population estimation using the multiple system estimator in the presence of continuous covariates, Statistical Modelling, 5, pp. 39-52.