# Imputation strategy in spatial data

Laura Martino and Alessandra Palmieri
European Commission – DG ESTAT
e-mail: alessandra.palmieri@ec.europa.eu

**Abstract**

In area frame surveys a mixed approach consisting in the observation of the surroundings of a not reachable point combined with orthophoto interpretation needs to be used in the locations that are particularly remote and difficult to reach. In this situation a simplified nomenclature has to be applied for some land cover categories due to the difficulties in properly distinguishing among specific classes. In the estimation phase the resulting observations can be considered affected by a sort of partial non response phenomenon. Classification of land cover indeed is available only at aggregated level. A donor based methodology is proposed to impute this missing detailed information. Assuming that neighbouring points are affected by spatial autocorrelation, potential sets of donors are identified among points within different distance thresholds. Capability of the method of correctly imputing the missing information is discussed and its robustness assessed in terms of two different cost-functions both based on the maximum distance observed among potential donors and recipient points.

**Keywords:** area frame survey, missing data, hot deck imputation, spatial data

## 1. Introduction

Area frame surveys usually foresee people going to the field and colleting in-situ information that are visible on the ground. This could be the case of crops, environmental parameters, forestry features and so on.
Since the accessibility to the point can be difficult for many reasons (fences, military areas, wild animals, etc.) it could be the case that for some units it is impossible to assess the land coverage in-situ. In these situations the recourse to a mixed approach - consisting in the observation of the surroundings of the point combined with orthophoto interpretation - is frequently adopted. As a consequence a simplified nomenclature needs to be applied for some land cover categories due to the difficulties in properly distinguishing among specific classes (i.e. durum wheat from oats and barley). In the estimation phase the resulting observations can be considered affected by a sort of partial non response phenomenon (some detailed information on land cover is missing). Classification of land cover indeed is available only at aggregated level.
Various methodologies are available to cope with partial missing data information (Little & Rubin, 1987). When data are affected by spatial correlation, the location of the sampling units can play an important role in the prediction of the missing information.

## 2. Imputation strategy for an area frame survey

One of the methodologies most commonly used to cope with missing data issues is the hot deck imputation (King C. S. & Bogle R. D., 2003, Gabriella Schoier, 1999). It consists of filling in missing values on incomplete records using values from similar, but complete records of the same dataset or external dataset. The identification of the best donor for each incomplete record can be based on different criteria like distance function matching or nearest neighbour. When spatial data are considered and sampling units are portion of land, physical distance among points usually represents a good indicator of similarity. Nonetheless to guarantee robustness of the imputation procedure, techniques taking into consideration the distribution of the set of donors need to be considered.

A methodology is proposed here taking into consideration both the need to look at the distribution of the land cover classes among the donor sets and the minimization of an overall indicator of distance between donor and recipient point.

The main steps of the methodology for each point affected by partial missing information are the following:

- Five distance thresholds are defined (10, 15, 20, 25 and 30 km);
- five nested sets of donors are set up composed of all the points belonging to the same stratum and lying progressively further off the recipient point (in a circle of ray equal to the threshold distance);
- a sort of 'standardized modal value' of the distribution of each set of donors is computed standardizing the relative frequency of each land cover class by the general share of the corresponding land cover in the country;
- the best donor set/value among those previously set up is selected maximizing a gain function;
- the modal value of the selected donor set is attributed to the recipient point.

### *2.1 The standardized modal value*

The standardized frequency of each land cover in a donor set is computed dividing the relative frequency of each land cover in each donor set by the corresponding share in the population

$$\hat{r}_{L,s} = r_{L,s} / r_L$$

Where

$\hat{r}_{L,s}$ standardized relative frequency of the Land Cover L-th in the donor set s-th

$r_{L,s}$ relative frequency of the Land Cover L-th in the donor set s-th

$r_L$ relative frequency of the Land Cover L-th in the population

The standardized modal value is the land cover class having the highest standardized relative frequency.

This device was introduced to avoid that the donor value was biased in favour of land cover classes that have the largest share in the general population.

### *2.2 The gain functions*

Two gain functions are proposed both linked to the maximum distance of the points belonging to each set of donors (ray of the circle) and the modal frequency of the land cover observed on the points belonging to each donor set. The aim of these functions is to favour the choice of the donor value most frequently found in the closest surroundings (measured as absolute distance or area) of the recipient point.

The first function is expressed as the ratio of the modal frequency and the area of the circle centred on the recipient point. It express how typical is the modal value in the area of circular shape surroundings the recipient point.

$$G_s = \hat{r}_{M_s} / \left( \left( Maxd_{M_s} \right)^2 * \pi \right)$$

The second cost function is based on the linear distance. It provides a measure on how further it is needed to go to find donor-value representative of the area.

$$H_s = \hat{r}_{M_s} / Maxd_{M_s}$$

Where:

k= $1,\ldots, s, \ldots, n_s$ set of donors satisfying the conditions

1) $\{ i \in s : i \in s+1, ....., n_s \}$

2) $\{ \forall i \in s, \; j \in s+1 : d_i < d_j \}$

$M_s$ modal land cover class of the distribution of the s-*th* set of donors;

$\hat{r}_{M_s}$ standardized frequency of the standardized modal land cover class of the distribution of the s-*th* set of donors;

$d_{M_s}$ distance of the donors having the modal land cover class from the recipient;

$Maxd_{M_s}$ maximum distance from the recipient point of the donors having the modal value.

## 3. A case-study: the European Land Use and Cover Area frame Survey (LUCAS)

The capability of the method of correctly input missing data is tested on the European 2009 LUCAS survey. The LUCAS (Land Use/Cover Statistical Area Frame Survey) survey is a field survey based on an area-frame sampling scheme (Martino & Fritz, 2008). Data on land cover and land use are collected and landscape photographs are taken. Eurostat carried out the largest ever LUCAS campaign in 2009. It collected data on the ground on land cover, land use and landscape diversity on approximately 234,000 points. Those points were selected from a standard 2 km grid with in total 1 million points all over the EU. The land cover and the visible land use data were classified according to the harmonized LUCAS land cover and land use nomenclatures.

The complete records of the LUCAS 2009 (http://epp.eurostat.ec.europa.eu/portal/page/portal/lucas/data/database ) have been used to test the capability of the method of properly imputing the missing data through a simulation exercise. Starting from the complete set of points with arable and permanent

crops in Europe (46,296 out of 234,907) the true land cover value of a single point has been deleted, one by one, and all the other points in the same country/stratum (arable land or permanent crop) have been treated as potential donors. Then distances between the recipient point and the others have been computed and five nested sets of donors defined according to the thresholds of 10, 15, 20, 25 and 30 km respectively. The new land cover category is imputed on the basis of the proposed methodology.
Some quality indicators have been computed to evaluate:

1. The capability of the methodology of imputing the correct value (unbiasness) by land cover (with 28 and 7 classes) and by country. This indicator is expressed as the percentage rate of agreement between the imputed and the true value;
2. The robustness of the obtained results with respects of the different distance thresholds and gain functions. This is expressed as the number of times the same land cover class is imputed out of the five potential sets of donors and with respect of the two different gain functions.

All the countries surveyed in 2009 have been included in the simulation. Their diversity in terms of land cover landscape (expressed as Shannon Evenness Index) has been accessed and analyzed in combination with the quality of the results to better understand whether it could be an important factor to improve the quality of the simulation.

The overall rate of agreement is not significantly different using the two gain functions ranging between 41% and 72% (depending on how detailed is the nomenclature adopted. See Table 1) for gain function 1 and between 42% and 73% for gain function 2. A large variability is observed at country level although.

Table 1: Overall rate of accordance with true land cover

| | Gain function 1 | | Gain function 2 | |
|---|---|---|---|---|
| | n. | % | n. | % |
| | Nomenclature 2 digit | | | |
| Disagreement | 31960 | 59 | 31600 | 58 |
| Agreement | 22286 | 41 | 22646 | 42 |
| | Nomenclature 1 digit | | | |
| Disagreement | 15138 | 28 | 14688 | 27 |
| Agreement | 39108 | 72 | 39558 | 73 |

The set of donor with the smallest size (10 km distance) seems to be the preferred one when using the gain function 1, while the largest set (30 km distance) is the one providing donation most frequently when it goes to the second gain function.

## References


Little R.J.A. & Rubin D.B. (1987) *Statistical analysis with missing data.* Wiley, New York.

Gabriella Schoier (1999) *On partial non response situations: the hot deck imputationMethod.* ISI99, Helsinki 10-18 August 1999, Finland

King C. S. & Bogle R. D. (2003) *Using Hot Deck Donor Imputation Methodology in the Service Annual Survey.* ASA 2003 Joint Statistical Meetings - Alexandria, VA, US

Martino L. & Fritz M. (2008) New insight into land cover and land use in Europe, *Statistics in Focus,* 33, Eurostat, Luxembourg