European Forum for Geostatistics Conference (EFGS)
The production & use of spatial statistics for sustainability
Statistics Portugal, Lisbon, 12-14 October 2011.

Agricultural holdings identified through the Italian 6° Agriculture General Census: methodologies adopted for their localization

Giampaola Bellini♥ - bellini@istat.it
Maria Grazia Magliocchi♠ - magliocchi@istat.it
Maria Antonietta Liguori♦ - liguori@istat.it

Abstract

The spatial aspect of the agricultural activity has always been investigated with great interest. In this context, as defined by the Regulation 1166/2008/Eu ruling the Sixth Agriculture General Census, agricultural and livestock holdings location will be realised not only at polygon level (the minimum administrative level, alias the municipality) but also at point level. Thus, with the geographical detail reached trough the 6° Census, it will be possible to refer agricultural holdings (AH) also to territorial classifications of environmental interest (i.e.: hydrological basins; Natura 2000 zones; National Parks) going beyond the traditional administrative level.

Processing the information collected through the questionnaire, namely the residence address, or the holding headquarter (HH) location, that is identified by the address and/or the cadastral map, will be possible to obtain the geographical coordinates referring to one information (address) or the other one (cadastral map). Depending on which information is used, two different approaches are possible.

The addresses can be processed with the Egon software, a program that validates addresses – giving also the possibility of coupling to the address the related geographical coordinates as geo-coded street graphs as Navteq and Teleatlas are included in its archive.

Moreover, the availability of cadastral maps makes possible to link each cadastral map code with the centroid (alias the polygon barycentre) coordinates of the considered polygon (map sheet or parcel).

Once a detailed location for each holding is defined, in order to avoid the identification of a single unit, the agricultural holding will be referred to the centroid of the census enumeration area (defined for population census purpose) in which the coordinates of the statistical unit previously identified fall.

This approach will determine the implementation of two final data sets, one related to the detailed coordinates – to be used for spatial analysis purpose - and the second one to be released to the broader public as diffusion coordinates. Choosing the census enumeration areas as reference polygon map, will help also in integrating information related to units identified through different censuses to a unique reference map. Nevertheless analysis based on agricultural holding punctually located will also be possible.

Each step of the adopted algorithm will be characterised by attributing a specific parameters to the identified coordinates, parameters that will be useful to describe the final geographical coordinates quality.

The detailed methodology set up to localize agricultural holdings, based on the mentioned tools available in Istat and on the constraint defined by law in order to avoid confidentiality disclosure, will be thoroughly described.

♥ Authored §§ 3; 3.1; 3.2 and 5.
♠ Authored §§ 2 and 4.
♦ Authored §§ 1 and 3.3.

1. Introduction

The spatial aspect of the agricultural activity has always been investigated with great interest. In this context, as defined by the Regulation 1166/2008/Eu ruling the Sixth Agriculture General Census, agricultural and livestock holdings location will be realised either at polygon level (the minimum administrative level, alias the municipality), and at point level. Thus, with the geographical detail reached trough the 6th Census, it will be possible to refer agricultural holdings also to territorial classifications of environmental interest (i.e.: hydrological basins; Natura 2000 zones; National Parks) going beyond the traditional administrative level. In this context, thanks to Geographic information systems (GIS), it will be possible either the integration of these data with much other statistical information and their dissemination to satisfy the increasing needs of new many data customers.

Many applications in this field are foreseen, as outlined also in the 'Geospatial Infrastructure Census Activities' of the United Nations, that can facilitate the analysis of phenomena for which the spatial dimension is relevant either they refer strictly to environmental aspects (relationship between agricultural production and drought and flooding events, impacts on vulnerable areas, etc.) or they are connected to social aspects (i.e. food data demand and food security issues).

Istat has oriented in the last decades its activity in geo-referenced statistical database production, management over time and release. Particularly Census output can be used in developing strategies of small area also thanks to the use of GIS application as census mapping database (an Istat application), that can be used in procedure for area mapping and data access via intranet-internet.

Other GIS applications at Istat are: (i) production of digital census atlas on CD-ROMs; (ii) geocoding of addresses on maps; (iii) visualisation and check of functional zones (i.e.: local labour areas); (iv) developing prototypes for data access via intranet and internet; (v) pilot study on morphological urban agglomerations.

2. Agricultural census questionnaire: questions to collect information on agricultural holding location

According to Eurostat, European Union Member States have to locate the agricultural holding headquarter (HH), releasing the geographical coordinates – latitude and longitude – to the nearest 5 minutes (in order to avoid the direct identification of ad individual holding)[1], which represent a land area of approximately 3,000-7,000 hectares, depending on the location in Europe.

In order to do this, with the 6th Agriculture Census, Istat has collected information about agriculture holding (AH) location that is – according to the mentioned Regulation – where main part or all agricultural production takes place. In Istat the definition adopted refer to the location "[…] where the building (one or more) connected to the agricultural activities is, within the agricultural land perimeter. This building can have different functions: it can be the holder residence or where the residence of agricultural labour force is, or the stable for livestock, or where mechanical equipment used for agricultural activity is stored, as well as buildings used for products storage purpose. Whether within the agricultural land perimeter there are no buildings, the holding headquarter is where the largest agricultural area is located[2]". Furthermore in case the localization of the holding

[1] The location of the agricultural holding is defined in Article 2(e) of the European Parliament and Council Regulation (EC) No 1166/2008. " 'Location of the holding' means the latitude and longitude coordinates within an arc of 5 minutes that avoid the direct identification of an individual holding. If a latitude and longitude location contains only one agricultural holding, then this holding shall be attributed to a neighbouring location, which contains at least one other agricultural holding", Cfr Regulation (EC) No 1166/2008 of the European Parliament and of the Council of 19 November 2008 on farm structure surveys and the survey on agricultural production methods and repealing Council Regulation (EEC) No 571/88, pag. 16.

[2] Cfr. "Istruzioni per la rilevazione – 6° Censimento Generale dell'Agricoltura", Istat, luglio 2010, pag. 40

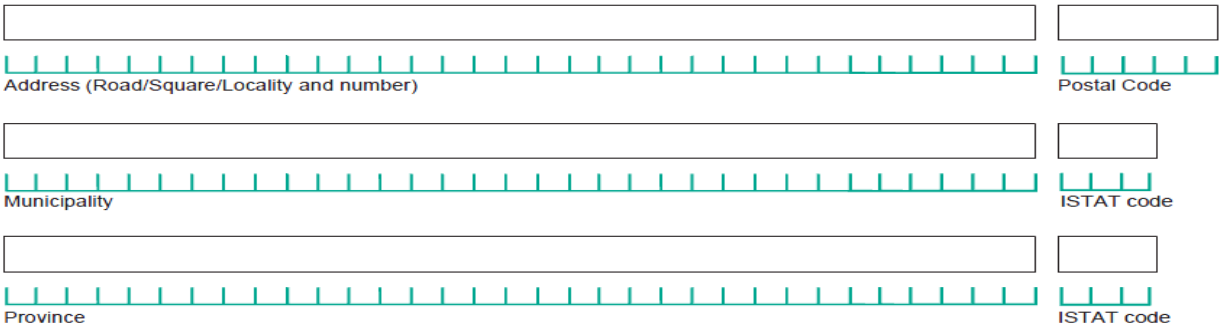
headquarter (HH) falls within 5 km from the holder's residence[3], it can be considered as the reference place of the AH.

Therefore, to identify the agriculture holding, information on the HH location (referring to the residence, to other buildings or to the main part of land) is necessary. This information has been collected through the agriculture census questionnaire, in two different sections, referring to the holder residence and to the other kind of HH, as described below. Supplementary information is required in order to establish if the HH lies within 5 km from the holder's residence.

Two main different approaches have been identified to locate HH: the address and/or the cadastral map. Using one information or the other one it will be possible to obtain the corresponding geographical coordinates.

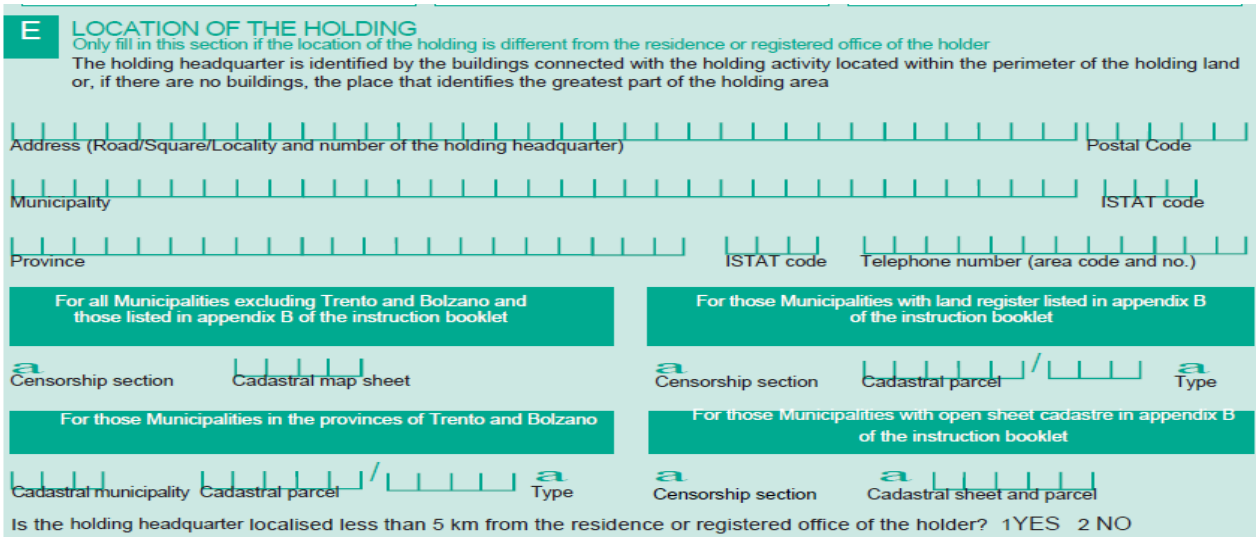
In the holder residence section (box A that is pre-printed with information derived from administrative archives and filled in another record by farmer if different) that is: address, municipality name and Istat code, province name and Istat code, postal code, telephone, e-mail and web-address (see Figure 1).

Figure 1 – Holder residence section



In the holding headquarter section (box E) in which, beyond the same information collected for residence location (but in this case is not pre-printed), it is required also the cadastral information as: cadastral section, cadastral map sheet and cadastral parcel, depending on kind of cadastre active in each specific territorial area. It has to be underlined that the holding headquarter section has to be filled in by farmer only if the HH location is different from the holder residence (see Figure 2).

Figure 2 – Holding headquarter location section



[3] "The holder's residence can be considered as the reference place only when it lies within 5 kilometres (in a straight line) of the place where the main part or all the holding's agricultural production takes place", Eurostat, "Handbook on implementing the FSS and SAPM definitions", Revision 7, 2007, pag. 7

Thus, addresses can be processed with the Egon software, a program that validates addresses – giving also the possibility of coupling to the address the related geographical coordinates as geo-coded street graphs as Navteq and Teleatlas are included in its archive. On the other hand, the availability of cadastral maps make possible to link each cadastral map code with the centroid (alias the polygon barycentre) coordinates of the considered polygon (map sheet or parcel).

In this way, for each HH, the coordinates of the centroid of the cadastral map, sheet or parcel, and/or of the holding address will be available and would be possible to use them at different territorial levels.

Nevertheless, in order to avoid the direct identification of each holding, Istat will provide to Eurostat, for everyone, the coordinates of the centroid of the census enumeration area in which the previously identified HH coordinates fall, so that the direct identification of an individual holding would not be possible.

3. Geo-referencing and geo-coding: an integrated approach

As mentioned above, GIS tools make possible to acquire and manage information for different territorial levels. In particular, as coordinates are known, the statistical information can be geo-referenced at point level, as such, or can be aggregated in functional zones by using GIS tools, reported in homogenous output zones or attributed through a spatial join to different polygon layer.

In the Italian case, Istat has decided to assign to each agricultural holding (or better to its headquarter) a specific census enumeration area and to release the related centroid coordinates to Eurostat. Furthermore the address matching procedure allows the geocoding of any statistical units having an address to the census enumeration area. The implementation of this kind of data sets will be of paramount relevance for territorial da analysis.

In order to assign geographical coordinates to any single AH, Istat adopted two different strategies: through address processing with a commercial software, named Egon, and through cadastral identification code matching.

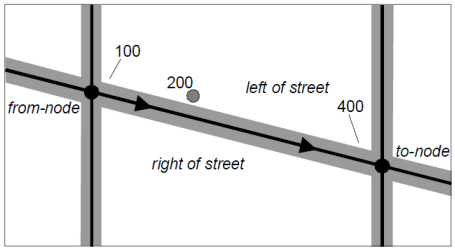
In both cases we have the possibility to merge a couple of coordinates to each location information.

3.1 Georeferencing through address matching

The recorded residence or HH address needs to be processed in order to be used for address matching purpose. First operation is the need to *normalise* the addresses, so that the street or road description is standardised. Moreover, also typing errors are corrected. In this way the address can be matched to the one included in the geo-referenced address archive. This database consists of street or road segments that are portions of street included between two intersections. In GIS, the streets are defined by lines representing the centre of the street, and the intersections are defined as nodes.

Thus, also an internal geographic attribute table, a from-node and a to-node, is defined for each line segment. The from-node and the to-node attribute determine the direction in which the line was digitized. Given the direction, it can be determined of which side of the street segment is left and which is right. Given this information it is possible by GIS tools to locate any given address on the street network in a process known as *address matching* (sometimes also referred to as *geocoding*). Each street address from a list is evaluated and matched to a location on the corresponding street segment, by an operation called interpolation, by which the location is chosen based on the address number in proportion to the range of addresses on the street segment (see Figure 3). This kind of approach requires that streets are numbered as described but it can be chosen for most purposes.

Figure 3 – Address matching



This process is internalised in the Egon software a commercial product that includes Navteq e Teleatlas georeferenced address data sets.

In the normalisation process, part is realised by Egon itself and part is realised by analysing the content of the agricultural holding record prior starting the data processing with Egon. An example of the normalisation process result is presented below (Table 1).

Table 1 – Address prior and post address normalisation process

Pre-normalised address	Normalised address
FRAZ. ROCHEFORT, 18	LOCALITA' ROCHEFORT,18
FRAZ. LIGNOD	FRAZIONE LIGNOD
FRAZ. LIGNOD-RUE TRACIASA, 45/3°	RUE TRACIASA,45/3
VIA S. GOTTARDO 4	VIA SAN GOTTARDO,4
VIA CASTELLO 1, FRAZ. POIA	VIA CASTELLO,1
Dorf 4, 39010 Kuens	DORF 4,39010 KUENS
P.ZZA G. B. LAMPI 2	PIAZZA GIOVANNI BATTISTA LAMPI,2
MEDEAZZA 6	LOCALITA' MEDEAZZA,6
Frazione Gramizzola	LOCALITA' GRAMIZZOLA
Frazione Cabosa	FRAZIONE CABOSA
VIA FONTE AVELLANA 140	VIA FONTE AVELLANA,140
Piazza Flora n. 1	PIAZZA GIUSEPPE FLORA,1

Regarding the *pre* Egon processing part, what needs to be implemented is the transformation of some specific GDZ (General Denominations Zoning) that are recognised by the software and the erasing of some undesired information recorded in the address field. For example if the municipality name follows the address description the normalisation procedure doesn't succeed.

Beside address matching function, Egon software also guarantees the generations of codes to identify the quality of the matching records. Four different values are defined: '0' and '1', that represent the geocoded units with different degree of precision (0 is the most and 1 is least precise) and '2', '3', that are the cases in which geocoding was not successful. Furthermore, the discarded records are identified according to different criteria adopted for exclusion. These cases are listed and coded and for each of them a specific solution might be found and applied through a second phase (interactive or automated) and, subsequently, through a second phase of data processing. Nevertheless, not all of them can be solved. The result of this process is a *point georeference*.

It has to be underlined that Egon software has been upgraded including also the list and codes of enumeration areas. A direct linkage between streets or streets numbers to enumeration areas is possible, otherwise a spatial join can be realised between address coordinates and the enumeration area polygon in which the address number coordinates fall. In both cases the centroid coordinates is the result of the join. The enumeration areas are the ones defined for the Census run in 2001 year and through a transposition table the linkage with the new enumeration area number is possible. In case the old 2001 enumeration area has been divided into 2 or more new 2011 enumeration areas

the new centroid coordinates should be identified by geographical overlay and provided, from cadastral map centroid. The result of this process is a *polygon geocoding*.

The census enumeration areas have been realised in the UTM-ED50 system, but the coordinates for this project have been released in WGS84.

3.2 Geocoding through cadastral maps

Prior using the digital cadastral data set for geocoding, the data set has been processed and data quality has been verified by Istat.

Referring to data processing, as data set has been provided at provincial level in the *cxf* format and in the Gauss-Boaga Roma40 reference system, except for the autonomous provinces of Trento and Bolzano-*Bozen* (managed at local level and for which the Datum is ETRS89). Maps have been converted into *shapefile* format and than the conversion into a different geographical reference system has been realised in order to have all the data into the same one. Only the relevant layers contained in the cadastral map have been considered (maps and parcels).

Cadastral units are organised by provinces and cadastral commune (relationship between cadastral commune and administrative municipality can be 1:1 or *n*:1 for several reasons).

In our Country, the cadastral maps have different characteristics in terms of typologies, projections, geometry and polygons encoding. For this reason a deep analyses was performed in order to understand the characteristics of each kind of cadastre.

At national level, the *Agenzia del Territorio* (AdT) is the authority with the task of maintenance of the central cadastral maps (except for Trento and Bolzano-*Bozen* provinces territories): encoding new cadastral commune and updating the centralised cartography with the one officially produced at local level.

Regarding projections in most cases the Cassini-Soldner is the one adopted at local level for map generation; referring to polygon units' hierarchy, in some areas the parcel is the smallest unit and each commune is divided into cadastral sheets and in turns these are divided into parcels. This is the most common case regarding the territories where the ordinary cadastre is active (referring to 7.023 administrative municipalities out of 8.094 – at January 2010 – see Table 2). In those cases the information referring to the map sheet level is enough detailed for AH geocoding purpose and respect the precision level required by Eurostat, as each map sheet can represent is below the threshold dimension defined by Eurostat.

Table 2 – Municipalities (at 1st January 2010) by cadastral typology in Italy

Cadastral typologies	N	%
Ex-Austro-Ungaric cadastre (BZ)	116	1.4
Ex-Austro-Ungaric cadastre (TN)	217	2.7
With open map sheet perimeter	690	8.5
Ordinary	7,023	86.8
Tavolare	48	0.6
Total	**8,094**	**100.0**

In other territories, cadastral map sheets (called with open perimeter) are spread among more than one municipality and in the territorial hierarchy the parcel represents the unit composing the municipality. In those cases the information referring to the parcel level was acquired. Regarding the dimension, each parcel can reach maximum 200 hectares.

In order to proceed with geocoding, centroid coordinates of cadastral maps or sheets, depending on cadastre type, have been extracted from the digital maps and joined with the administrative units in which they fall.

Referring to data quality, information on completeness (partial or complete lack of map sheets by municipality) and geo-referenced data quality (special calibration realised or not on ortophotos) has been provided by *Agenzia del Territorio*.

Nevertheless, data check was performed by Istat and referred to:

a. completeness in terms of municipalities mapped (for 305 municipalities cadastral maps are not available at Istat); completeness in terms of cadastral units (sheets or parcels) cannot be assessed as the complete map sheets or parcels list by municipality is not available and data check GIS through GIS would be strongly time and resource consuming;

b. logical accuracy (1st level): an overlay operation made possible to check centroid falling out of municipality boundaries (and thus out of national land) and between enumeration area polygons (0.6% out of total involved units);

c. logical accuracy (2nd level): check the list of the cadastral commune spatially joined administrative with the list released by *Agenzia del Territorio* in order to identify cases in which the cadastral map coordinates do not fall in the correct administrative unit.

Extracted centroid coordinates are spatially joined to census enumeration areas, so that we have their centroid coordinates coupled to each cadastral polygon.

Than a unique key code is created and another matching procedure is applied. In doing so, the enumeration area centroid coordinates can be coupled to each specific HH.

Also this procedure will produce some discarded records, which characteristics are analysed to assess feasibility of correction procedure possible only where typing errors are identified.

At present the following frequencies have been obtained for a sub-set of information collected through the specific section E of the questionnaire on the agricultural holdings headquarter location.

Table 3 – Information collected on HH by cadastral typologies
(provisional data)

Cadastral typologies	N	%
Ordinary	414,299	37.2
'Tavolare'	152	0.0
Ex-Austro-Ungaric cadastre (TZ e BZ)	12,034	1.1
With open map sheet perimeter	1,937	0.2
Without cadastre insert or discarded	684,574	61.5
Total	**1,112,996**	**100.0**

Source: 6th Italian Agriculture General Census

3.3 Assigning geographic coordinates to residues

Other assigning procedures have been identified for the imputation of coordinates to all that units which couldn't be geo-referenced through address matching or through geo-coding cadastral maps. In particular, as first option the information, if filled by the holder, on the 'distance of less than 5 km' will be used. In other words, if there's the information on the 'distance of less than 5 km', to all these units will be assigned the coordinates achieved through residence address matching, under the hypothesis that the holder's residence can be considered as the reference place of the AH only if located in the same municipality.

As further step, all the residues will be treated in the following way. Through the Istat agriculture holding identification code, these records will be linked to that one present in the pre-census list and the information on holding headquarter estimated for those will be used.

Finally, all observations for which it will not be possible to link a pair of coordinates, will be further analyzed.

The objective of this analysis is to assign the missing value (the geographical coordinates) through an allocation method. In fact there are different allocation methods – deterministic or stochastic – for partial missing value, but all are implicitly or explicitly based on the assumption that missing data are missing at random (MAR): so the probability of having, for a given variable, a non-response does not depend on the value of the non-response itself.

The general characteristic of the deterministic methods is that the repeated imputations for units with the same characteristics produce always the same imputed values.

The stochastic methods are characterized by a random component corresponding to a probabilistic scheme associated with the selected method of allocation and then the assignment procedure - for units with the same characteristics - could produce different imputed values.

In both methods, before identifying the value to be attributed, it is better to bring back data to homogeneous subsets, cells, or classes of imputation, obtained through a series of appropriate layers.

After the census data analysis, among the deterministic methods, we suppose to use that of the nearest-neighbor while among the stochastic methods that of the hot-deck.

4. Final data quality and metadata production

As geographical statistical data are produced a metadata production/description is necessary.

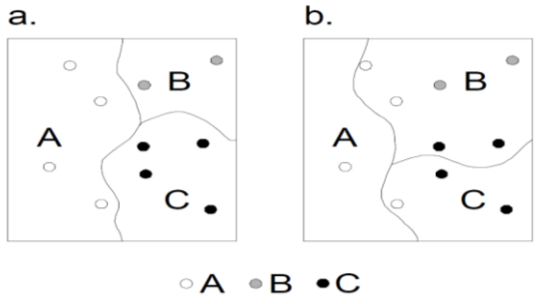
Referring to the issues to be taken into account in producing metadata the main points are: i) logical accuracy; ii) positional accuracy; iii) semantic accuracy; iv) temporal information; v) data set completeness.

Particularly, logical accuracy refers to the integrity of relationships among geographic features. A town represented as a point in one GIS database should fall into its corresponding administrative unit in another GIS layer.

Figure on the left, for instance, shows a set of sample survey sites that have been determined, using a very accurate global positioning system.

The base map in Figure 4 on the right, in contrast, while logically accurate, has a low degree of positional accuracy. Some of the accurately measured GPS points, therefore, fall into the wrong administrative units.

Figure 4 – Logical accuracy



Positional accuracy, in contrast, maintains that the coordinates of features in the GIS database are correct relative to their true positions on the earth's surface.

All the steps realised in generating the HH geographical coordinates data set and information about data quality, formats, processing have the role of facilitating the maintenance and updates of digital data sets held by the producer, of releasing information for an appropriate use by data users, of increasing the correct integration of different data set.

For this reason all the steps involved in data production will be properly documented.

Referring to logical accuracy, the geographical coordinates of each holding have to be located in the administrative unit in which the HH is declared by the farmer in the questionnaire.

Referring to positional accuracy European regulation requires that the HH geographical coordinates position shouldn't be farther from real position on earth than 2,5', meaning that AH have to be located in an arch of 5', which corresponds to 10 km at our latitude.

The assessment of the precision of the data set obtained is not directly measurable, but in the steps realised we can document the approaches taken, through specific parameters attributed to each couple of coordinates.

Referring to the geo-referencing procedure, the first parameter documents the information used for geo-referencing purpose. The values vary, depending on how coordinates are obtained:

1 by address matching;

2 by address matching, where the address is the one of the holder residence that is within 5 km from HH and in the same municipality;

3 by cadastral identification code matching;

4 without geographical coordinates.

Furthermore, a second parameter is defined to document the quality of coordinates released to Eurostat that can be:

1 for enumeration area coordinates generated through spatial join of the first two previous cases;

2 for enumeration area coordinates generated through spatial join of the third previous case;

3 for enumeration area coordinates generated through data imputation.

Furthermore, other two parameters have defined to express the '*fragmentation*' of the AH. In fact, the AH is generally composed of land that can be located in a single or more municipalities and in one or more land units, so the identification of the single HH will be less representative of the total physical AH. Thus a specific parameter refers to number of municipalities on which the AH manage its land or livestock and another to the land unit number composing the AH.

5. Conclusion

In order to meet European Union requirements Istat will locate agricultural HH to the centroid coordinates of the census enumeration areas in which they fall. Furthermore a more detailed geographical coordinate will be available for each AH, depending on kind of information collected on location of its HH by the agriculture census questionnaire (either address or cadastral map).

All the geo-referencing and geo-coding process has been defined and implemented in order to meet precision requirements defined by Eurostat and geographic coordinates will be released with a set of metadata describing their quality.

The availability of geographical coordinates for each AH will allow - through the exploitation of GIS tools - the possibility to refer agricultural holdings also to territorial classifications of environmental interest (i.e.: hydrological basins; Natura 2000 zones; National Parks) going beyond the traditional administrative level.

As enumeration areas are also the basic polygon map for representing all surveyed phenomena by censuses, this approach will potentially allow the aggregation of data collected at this detailed territorial level.

Furthermore the address matching procedure allows the geocoding of any statistical unit having an address to the census enumeration area in which falls.

Proceeding in this direction, it'll be possible to perform more detailed and relevant territorial analysis.

References

1. Bellini, G. Istat. *Rapporto del Gruppo di Lavoro ISTAT per l'analisi di metodi, strumenti e tecniche di georeferenziazione delle aziende agricole da applicare nel contesto del 6° Censimento Generale dell'Agricoltura (tempi, tecnica di rilevazione e strumenti utilizzabili).* Istat. Roma, Luglio 2009.
2. Crescenzi F., Fortini M., Gallo G. e Mancini A., Istat, *La progettazione dei censimenti generali 2010 – 2011 6 - Linee generali di impostazione metodologica, tecnica e organizzativa del 15° Censimento generale della popolazione,* Documenti Istat, No. 6, 2009
3. Crescenzi F., *La georeferenziazione dei terreni delle aziende agricole nella prospettiva censuaria*, http://www3.istat.it/istat/eventi/2007/interconferenza/materiale/crescenzi.pdf, 2006
4. European Parliament and Council. *Regulation (EC) No 1166/2008 of 19 November 2008 on farm structure surveys and the survey on agricultural production methods and repealing Council Regulation* (EEC) No 571/88.
5. Eurostat Working Group "Structure of Agricultural Holdings". *Handbook on implementing the FSS and SAPM definitions – revision 7, 2010*. CPSA/SB/652 rev. 7.
6. FAO, Statistical Development. *A system of integrated agricultural censuses and surveys, World Programme for the Census of Agriculture 2010.* Volume 1, Series, N.11.
7. Grande E. & Luzi O. *Metodologie per l'imputazione delle mancate risposte parziali: analisi critica e soluzioni disponibili in Istat* in Contributi Istat, 6/2003.
8. Istat. *Istruzioni per la rilevazione – 6° Censimento Generale dell'Agricoltura.* Roma, Luglio 2010.
9. United Nations, Department of Economic and Social Affairs Statistics Division. *Handbook on Geospatial Infrastructure in Support of Census Activities*. Studies in Methods. Series F No. 103, New York, 2009
10. United Nations, *Censuses, geocoded data and small area statistics,* Supporting paper submitted by Istat, (edited by) Crescenzi F, CES/2001/21/Add.4, 4 May 2001
11. United Nations, Department of Economic and Social Affairs Statistics Division. *Handbook on geographic information systems and digital mapping*". Studies in Methods. Series F No. 79, New York, 2000.