Agricultural administrative sources data quality: a proposal for standardized indicators

Cusimano, S.[1], Bruni, A.[2], Fusco, D.[3]

[1] *Istat, Rome, Italy; cusimano@istat.it*
[2] *Istat, Rome, Italy; anbruni@istat.it*
[3] *Istat, Rome, Italy; dafusco@istat.it*

Abstract
In the field of agricultural statistics, Istat focus has shifted on the build up of a register of units operating in the primary sector. The realization of the Farm Register provides a basic framework widely applicable. Administrative and statistical sources are integrated to gather the different kind of information on the agricultural domain: the main administrative sources handled are the Integrated Administration and Control System (ICAS) and the System for the identification and registration of Bovine animals. Micro-integration and its processing procedures are made difficult by the nature of the sector, whose peculiarities make complex the correct identification of units as well as the estimation of their actual size and their principal activity. With the extensive use of administrative sources, it is of vital importance to determine the statistical usability of the farm register on a regular basis. It means to determine the key quality constituents of administrative data and the derived register in a systematic, objective, and standardized way. So, the aim of the paper is to identify the best fitting quality indicator available in relation to agricultural administrative sources and variables.

Keywords: (1-5 words), farm register, administrative data sources, quality assessment, quality indicators.

1. Introduction

Fundamental transformations are affecting the environment in which producers of official statistics are operating: the informative context for many National Statistics Institutes (NSIs) has noticeably transformed since administrative data have turned exploitable in a wide array of application, i.e., in order to create statistical registers. Administrative data are the major input to these registers, with their well-known strengths and weaknesses.

A modernization scheme has been revisiting Italian NSI business model to foster the ability to respond to emerging needs: the relevance to invest in a register-based approach to

industrialization has definitely emerged. Following the new scenario, Italian NSI is now entering the number of countries that manage a Statistical Farm Register (SFR) moving therefore from an ordinary census-frame procedure to an up-to-date streamlined infrastructure. Comprehensive and timely information from different sources should be reliable and coherent: the setting up of a register of farm and farm holdings is the key pillar needed in order for agricultural statistics to be "fit for the future", telling the story of Italian agriculture and providing a cost-effective way of producing unique information used for decision-making in private and public sectors, while reducing the burden on respondents.

When a sample survey is performed, data quality ultimately depends on non-sampling errors (i.e., skills and experience of interviewers, effectiveness of measurement instruments), as the sampling error has been given by fixing sample size. When statistics are produced on the basis of administrative data sources, data have already been collected so the accuracy of microdata is determined by dissimilar features. Then, as stated in Daas *et al* (2009), the statistical usability of a data source requires to be examined by an NSI prior to its use.

In recent years, quality frameworks have been enhanced to assess the quality of administrative data sources in an efficient and standardized way, to determine if an administrative source can be actually used for statistical purpose and how. The field of general principles and guidelines is developing quickly: a recommended suite of quality measures whenever administrative data are used needs to be outlined. So, the aim of the present paper is to identify the best fitting indicator available to assess (i) input quality, at microdata level, of available agricultural administrative sources and (ii) output quality of the Italian SFR. The quality assessment will be done on a prototype SFR version.

2. A case study: the Italian Statistical Farm Register.

A SFR represents a key element in the production of agricultural statistics by providing survey frames (with stratification e.g. by size, type and location) and by contributing to the

integration of agricultural issues with that of other sectors. Many sources are matched when a statistical register is created. In Italy, the SFR build-up consists of integrating information coming from six administrative sources (Integrated Administration and Control System, Animal register, Tax declaration on agricultural land, Land cadaster, Chambers of Commerce, Value Added Tax - VAT on agricultural income) and four statistical sources owned by Italian NSI (Business Register, Agricultural Census, Survey on rural tourism accommodations, Survey on quality products 'Protected designation of Origin' – PDO, 'Protected Geographical Indication' – PGI and 'Traditional Speciality Guaranted' – TSG).

The capture process, processing and integration of administrative sources referring to the agriculture sector has specific difficulties to face, due to the peculiarities of the sector and the sources. The agricultural sector is characterized by small and very small farms, with family labor force. So, it complicates the correct identification of the target statistical units as well as their size estimation and principal economic activity.

Regarding administrative sources, there is the lack of a unique archive deemed as pivotal: the main administrative source used is the Integrated Administration and Control System (IACS). In term of coverage, IACS declarations identify all farms receiving economic subsides, but farms with crops not covered by Community Agricultural Policy are absent. Moreover, the sector information is collected for different purposes (e.g. in IACS the extent is the contribution payments, so that applicants for IACS are not necessarily holders; in the Animal register, the scope is the registration of animals for public health reasons).

The following table summarizes the acquisition of administrative data in SFR by outlining the relevance by source:

ADMINISTRATIVE SOURCE	**RELEVANCE**
IACS	Identify farms, localization and structural characteristics
Animal register	Identify farms with livestock and the number of animals, considering species
Tax declaration on	Coverage of "potential" farms: it includes persons, simple partnership and non-

agricultural land	commercial entities
Chambers of Commerce	Identify companies with agricultural economic activities, primary or secondary
VAT on agricultural income	Identify companies with VAT declared in agricultural activities
Land cadaster	Contains information about land parcel owned by the same person, having the same quality or class and the same assigned use

Walgreen and Walgreen (2014) provide an extensive description of register-based statistics: complete listing and known identities are set as the building blocks of a register. Then, each administrative dataset requires a preprocessing treatment to identify the farm manager personal code and the land quality according to IACS and Istat classification. After the preprocessing phase, the deterministic micro-integration of all sources takes place using personal code as the matching key.

The integration of several data sources increases the possible conflicts into the available information (Unece 2015b), so a hierarchy to the different sources is given, the units being put into four lists according to the sources of origin, namely:

LIST 1 IACS and Animal register (AR);

LIST 2 Tax declaration on agricultural land and Land cadaster (TD and LC);

LIST 3 Chambers of Commerce and VAT on agricultural income (CC and VAT);

LIST 4 Survey on rural tourism accommodations, Survey on quality products PDO PGI and TSG, Agricultural Census.

After the linkage of the lists, based on the hierarchical lists structure, it is been built a set of eligibility rules in order to finalize the composition of SFR. The resulting output is organized in three databases, linkable by personal code:

- Farms and main structural characteristics (e.g. Utilized Agricultural Area, main crops, livestock, eligibility rules, etc.);

- Farms and personal data (e.g. Municipality, Region, Address, telephone number, etc.);
- Land parcels (e.g. crops quality, surface, localization, etc.).

3. Quality assessment

Considering the objectives previously outlined and the available resources, among the indicators developed by international literature (Zabala *et al.*, 2013) the most significant in relation to agricultural administrative sources and variables were selected.

The quality assessment should start by drawing up the documentation of the quality aspects, in order to make a decision regarding the fitness-for-use of the administrative data source. The quality evaluation can been borrowed from Statistics Netherlands hierarchical approach, where the nested application of *Metadata* and *Source* hyperdimension constitute the Discovery Phase of the data acquisition process (more details can be found in Daas *et al.*, 2011). The *Source* hyperdimension measures the extent to which information contained in a data source is exploited. The *Metadata* hyperdimension focuses on the conceptual and process related quality aspects of the source metadata. The results shown were obtained by comparing the evaluation results for every measurement method for each quality indicator in each dimension and selecting the most commonly observed score. The symbols for the scores used in table 1 and 2 are: good (+), reasonable (o), poor (-) and unclear (?).

Table 1 - Evaluation results for source, *Source* hyperdimension

DIMENSION	**DATA SOURCES**					
	IACS	**AR**	**TD**	**CC**	**VAT**	**LC**
Supplier	+	+	+	+	+	+
Relevance	+	+	o	-	-	o / -
Privacy and security	+	+	+	+	+	+
Delivery	+	+	o / -	+	+	-
Procedures	o / +	o / +	o	-	o / -	-

Table 2 - Evaluation results for source, *Metadata* hyperdimension

DIMENSION	DATA SOURCES					
	IACS	**AR**	**TD**	**CC**	**VAT**	**LC**
Clarity	+	+	+	+	+	+
Comparability	o / +	-	-	-	-	-
Unique keys	+	-	+	+	+	o / -
Data treatment	o / -	o / +	-	-	-	-

IACS *Integrated Administration and Control System*
AR *Animal register*
TD *Tax declaration on agricultural land*
CC *Chambers of Commerce*
VAT *Value Added Tax on agricultural income*
LC *Land Cadaster*

In the evaluation at the *Source* level (table 1), the low scores for all sources on the *Procedures* dimension depend on problems at the fallback scenario indicators, caused by the absence of emergency measures when data sources are not delivered according to arrangements made. The result of Metadata evaluation are shown in table 2. In dimension *Data treatment* the negative outcome comes because of the high number of checks and modifications of the data by the data source keeper.

Considering the main issues connected the specific field (e.g. the absence of a unique complete administrative source, the absence of a unique farm definition, etc.) the overall result can be considered as satisfying. Furthermore, the choice of hierarchy given to the sources in the processing procedures is confirmed by the checklist framework.

At the second step, the quality assessment is focused on the output evaluation according to the standard dimensions of EU quality vector. For example, users want SFR data to be accurate and timely. Accurate means that the information recorded portray reality. Timely implies that the data are released in a punctual manner, with time lag as short as possible.

According to the most recent literature, the output quality assessment measure has applied at four dimension: timeless, accuracy, accessibility and coherence, using indicators shown in table 3.

Table 3 – Output quality measure for SFR

Dimension	**Indicators**	**Value**
Timeless	Temporal lag, measured in months, between the dissemination date of SFR and the reference year to which they refer	Critical
Accuracy	Coverage:	
	- Number and % of enterprises active in t and as change from t-1	Satisfactory
	- Number and % of units by list	Satisfactory
	Completeness:	
	- Farm Company name: Number of units and % with company name missing	Fair
	- Address: Number of units with address missing	Fair
	- Telephone: Number of units and % with telephone number (fax, email) missing	Critical
	Number and % of questionnaires rejected by type of error	N. A.
	Number and % of units with wrong address	N. A.
	Measurement errors	N. A.
Accessibility	How many readily available relevant data are to the users	N. A.
Coherence	Coherence with other surveys	Satisfactory

Each dimension is calculated considering the prototype version of SFR (year 2013) and the indicators are chosen to be suitable at the variables available in the SFR. The results of indicators released in percentage are evaluated considering the following ranges:

- Critical ≤ 70;
- $70>$ Fair ≤ 90;
- Satisfactory > 90.

For the other ones:

- 1 Satisfactory
- 0 Critical.

Timeless is calculated considering the release expected time. A major effort will have to be made in relation to achieve an higher timeless and the next versions of SFR are planned to be released after two years from the base year indeed.

About coverage, the indicator number of enterprises active in t and as change from (t-1) is calculated considering VI Agricultural Census data. For the later versions of SFR it will give a basis for comparing the different versions of the register, e.g. the difference between SFR 2014 and SFR 2013.

The not satisfactory values of the identifying characteristics depend on the evaluation of the prototype version. Also measurement errors, wrong addresses and questionnaires rejected will be calculate after the testing phase has ended. It was not possible to calculate the accessibility indicators because, indeed, the SFR has not yet available to users.

Coherence is calculated considering Farm structure survey 2013 (FSS) also known as Survey on the structure of agricultural holdings (Reg. EC 1166/2008). The satisfactory score is assigned by comparing the number of farm and the Utilized Agricultural Area by regional territorial breakdown.

4. Final remarks

The present paper should be seen by a perspective of sharing of experience concerning register-statistical methodology and quality issues. The results described show that the quality framework developed for administrative registers is a suitable tool for the evaluation of the SFR quality. The main aim to determine a systematic and standardized way to evaluate SFR input and output has been achieved.

As a rule, the understanding of how an administrative source should best be used by a statistical office requires time to develop. The coverage of the population of farms has to be carefully analyzed. Before implementing the final methodology, the goodness of the complete list is going to be tested through a quality survey to detect and correct frame errors. Then, a specific survey has been launched beginning last April and outcomes are awaited by end of the year: these results will help to improve processing procedures and to adjust, when necessary, eligibility rules in order to obtain the final version of the SFR in its first edition.

Quality improvement is an iterative process (Unece 2015a): indicators based on temporal consistency between two or more consecutive SFR edition could provide further evidence on quality e.g. farm births and deaths tracking viewed as a time series. Further deepening is also needed in describing the conditions for applicability of the methods proposed in the literature.

5. References

Daas, P., Ossen, S., Vis-Visschers, R., & Arends-Toth, J. (2009), Checklist for the Quality evaluation of Administrative Data Sources. Statistics Netherlands, The Hague/Heerlen

Daas P., Ossen S. (2011). Report on methods preferred for the quality indicators of administrative data sources, Blue – ETS Project, Deliverable 4.2.

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.

Wallgren A. and Wallgren B. (2014), Register-based Statistics: Administrative Data for Statistical Purposes, J.Wiley & Sons 2nd edition, Chichester, U.K.

Unece (2015a), Guidelines on Statistical Business Registers, http://www.unece.org:8080/fileadmin/DAM/stats/publications/2015/ECE_CES_39_WEB.pdf

Unece (2015b), Statistical Network Methodologies for an integrated use of administrative data in the statistical process - Administrative data (MIAD), Deliverable A.2. Different contexts for the statistical use of administrative data https://ec.europa.eu/eurostat/cros/sites/crosportal/files/Different%20contexts%20for%20the%20statistical%20use%20of%20administra.pdf

Zabala, F., Reid, G., Gudgeon, J. and Feyen, M. (2013), Quality Measures for Statistical Outputs using Administrative Data, Statistical Methods, Statistics New Zealand.