An integrated data collection system for farm structure statistics

**Keywords:** Integrated data collections, agricultural census, validation, dissemination

# Introduction

In 2020 the farm structure data collection, which is run as an agricultural census, covers more than 400 variables for around 10 million farms in the EU. This data covers 98% of the utilised agricultural area and livestock units. The cost of this exercise adds up to 281 million euro and the burden of providing data is perceived as considerable according to the European Agricultural Statistical System. Until 2016 data collection, the data processing at Eurostat relied on a traditional process, which depended largely on a series of non-automated procedures.

For the 2020 agricultural census, Eurostat is carrying out the Eurofarm 2020 project, with the main objective of modernising the statistical production system. An integrated approach between different data collections, an automated system for data processing and the modernisation of the data collection methods are expected to result in a reduction of the cost and an increase in data quality.

# Methods

From the technical point of view (legal and other aspects will not be analysed in this abstract), the project for the modernisation of the farm structure statistics followed a dual approach:

* More efficient and technologically advanced data collection methods were fostered by Eurostat and the national data providers.
* More automated, efficient and integrated data production system, which integrates several data collections, follows a modular approach and replaces data processing tasks, which involved a high degree of human intervention, by an automated system designed, developed and implemented by Eurostat.

Additionally, Eurostat has worked to improve the dissemination methods in order to better satisfy user needs.

Below is an overview of the methods applied to carry out this modernisation project in the context of the three relevant phases of the statistical business process model.

## Collect

Data collection for farm structure statistics was based until 2016 on diverse methods such as the computer assisted personal / telephone / web interviewing, postal data collection, pencil and paper interviewing, use of administrative registers or a combination of them. For the 2020 data collection, many countries take into use new data collection methods:

* Administrative registers: Eurostat has collaborated with NSI to access data from registers available under the common agriculture policy and overcome the practical challenges that this entails. This method reduces simultaneously response burdens for farmers and costs to national data providers.
* Big data sources and innovative approaches: Data collected from remote sensing (earth observation with satellites) are gaining importance.

For the microdata collection, a modular methodology is used, which links core data, module data and, when the farms above the threshold do not provide enough coverage, frame extension. The core covers general, land and livestock variables while the modules cover labour force, rural development, animal housing, irrigation, soil management, machinery and equipment, orchard and vineyards.

## Process

Integrated Farm Statistics (IFS) consists of eight different data collections: farm structure microdata, standard output coefficients, use of administrative data sources, non-significant or non-existent variables, relative standard errors, annual working unit factors, data collection thresholds and financial information. Eurostat has set up a data production system, which validates automatically each of those data collections and integrates them in order to obtain outputs that satisfy user needs.

## Disseminate

The Eurofarm 2020 project allows for a more timely publication of the statistical tables and an improved dissemination plan including new items such as the publication of historical series and the release of scientific use files. The former requires the adjustment of the data collection thresholds in order to have data comparable over time and within countries. For the latter, Eurostat has developed a methodology to disseminate microdata at farm level for researchers while ensuring the respect of data confidentiality [1].

# Results

The main results are presented below.

## Collect

The modular strategy and raised thresholds are estimated to account for a reduction of the cost of 56 million euro as compared to the previous census and a similar proportion is expected for the 2023 and 2026 data collections. This adds to the reduced statistical burden that stems from the use of lower precision requirements for module variables.

Currently, there are no specific estimations on the savings that stem from the use of new data sources but they have significantly reduced the cost and the statistical burden. The administrative data sources have the highest impact. Their use has spread in the context of this modernisation project. As shown in table 1, the number of countries, which collect more than half of the variables from administrative sources, has increased considerably.

Table 1. Number of countries by % of variables collected from administrative data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Less 25% | 25 to 50% | 50 to 75% | More 75% |
| 2020 | 12 | 4 | 9 | 3 |
| 2016 | 12 | 10 | 6 | 0 |

## Process

Under the Eurofarm 2020 project, the above-mentioned data collections are processed automatically. This data processing entails the integration of all the 8 data collections, the validation at all levels, new variables are derived after crossing information from these datasets, EU aggregates are computed and the files to be published are produced after treatment for confidentiality and reliability.

Figure 1. Validation levels according to the European Statistical System [reference]

Because of the complexity of integrated farm statistics (over 10 million records with hundreds of variables and 8 different datasets), setting up an automatized validation process was very challenging. It was necessary to define and develop hundreds of rules and test them by using thousands of test files. Below is an overview of the resulting validation procedure, which is, performed at all levels (figure 1 summarizes the different validation levels):

* Validation Level 0 “Consistency with the expected IT structural requirements”: a fully automatic procedure checks the structure of the dataset to avoid typos, basic errors and ensures completeness and correctness of code-lists.
* Validation Level 1 “Consistency within the dataset”: a fully automatic procedure checks that the contents of the dataset are self-consistent. It contains more than 500 rules to different degrees of complexity.
* Validation Level 2 “Consistency with other datasets within the same domain and data source”: a mostly automatic procedure where the system computes different ratios and statistics. Then it triggers warnings for values, which are potential outliers. An automatically generated report is sent to the data senders to confirm and explain each of the potential outliers. Among the checks performed at this level: time series analysis for each quantitative variable at the lowest possible level of disaggregation, values out of confidence intervals for cross – dataset ratios and fully automated validation rules which use data from different datasets.
* Validation Level 3: “consistency within the same domain between different data sources”: a mostly automatic procedure checks if different data senders, which are expected to have similar developments in certain variables, provide consistent data. “Mirror checks” are not relevant for this domain.
* Validation Level 4: “Consistency between separate domains available in the same organization”: a mostly automatic procedure executes a cross-domain comparison. The system automatically retrieves data from other domains and compare them to the dataset to-be-validated; values falling out of a confidence interval are reported. The data sender has to confirm and provide an explanation. Since the data from different domains are not directly comparable, the system is prepared to make the appropriate transformations.
* Validation Level 5: “consistency with data available in other organisations”: a fully automated procedure (at Eurostat) for the transmission of a subset of one of the data collections to another organisation, which performs, further checks.

## Disseminate

The efficiency gains in data processing will increase the timeliness in the publication of statistical tables. Additionally, two new dissemination products are designed:

* Farm structure time series: Currently farm structure data go only as far back as 2005. Moreover, the figures are not fully comparable over time as there are break in the series. The historical data are to be adjusted from data collection thresholds so that comparable time series are available as far back as 1990.
* Scientific use files: Anonymised microdata available for research institutions for specific research projects. The production system applies a series of methods to blur the data in a way that individual farms cannot identified: Micro-aggregation, rounding, merging or classifying variables in percentage bands.

# Conclusions

In the context of the agricultural census 2020, Eurostat and the European Statistical System carried out a project to better integrate and validate the farm structure data collections. This reduced the cost for the taxpayers, decreased the statistical burden to the farmers and improved the quality of the data. This took place in a wider context of modernising agricultural statistics and affected several steps in the statistical business process model. This has been achieved by:

* Replacing traditional data sources by using administrative registers, modernising the data collection methods and switching to a modular approach with increased thresholds.
* Setting up an automated data production system with less human intervention and a much higher number of validation checks with higher complexity.

The improved timeliness and more data made available is expected to produce a considerable gain to the users of farm structure statistics.

# References

1. Ramos H et al, “Disseminating European data at farm level while keeping confidentiality: scientific use files and homogeneous spatial units”, ICAS VIII (2020): <http://icas2019.icar.gov.in/pdf/Abstracts%20for%20Thematic%20Oral%20Presentations.pdf>