

An innovative approach to using a variety of data sources and statistical methods for micro-level statistical collection

Anders Grönvall, Swedish Board of Agriculture, anders.gronvall@jordbruksverket.se

Ylva Olsson, Swedish Board of Agriculture, ylva.olsson@jordbruksverket.se

Jesper Fransson, Swedish Board of Agriculture, jesper.fransson@jordbruksverket.se

Abstract

By using various sources, we have been able to produce statistics of sufficient quality and with a reduced response burden.

In the framework of agricultural statistics, the most important survey is the Farm Structural Survey. The EU surveys of agricultural structure are currently regulated by Regulation (EU) 2018/1091, which specifies, among other things, what data must be collected and how to collect them. The regulation states that Member States shall use one or more of the following sources or methods to obtain micro - level data:

- *statistical surveys;*
- *administrative sources (as specified in the regulation);*
- *other sources, methods or innovative approaches.*

The third option was new for this regulation, and the guidelines did not give specific guidance on how to implement it.

During the preparatory work for the 2020 census, we evaluated different possibilities for collecting data for each variable. One part of the survey covered animal housing and manure management and comprised 70 variables. Earlier farm - structure surveys in the 2010s also included variables on animal housing and manure management. From these surveys, we learned that such variables are difficult to collect with high quality through questionnaires.

The preparatory work showed that no single administrative register could cover the entire information need for these aspects. However, a substantial amount of relevant data exists in different places: administrative sources, information from advisory organisations, other statistical sample surveys, legislation, etc. Our solution was to combine all these data sources, link them to our frame of holdings, and use statistical methods to generate micro - level data for each holding. The large number of variables posed a major challenge, requiring a wide range of methods. For some variables a single method was

sufficient; for almost all variables several methods and sources were applied. The experience from this project will have a significant impact on future statistical production.

Keywords: innovative approach, multi-source statistics, agriculture surveys, administrative data, data integration

1. Introduction

Sweden have conducted statistical surveys of the structure of agriculture for almost a hundred years. Until the end of the 20th century Sweden only used questionnaires that were sent to the agriculture holding. Since 2000 Sweden have used a combination of regular statistical survey and linking data from administrative registers.

Since 2000, the possibility of linking register data has been continuously evaluated for each variable in the survey. Due to decreasing response rates, and since 2020 the possibility to use “other sources, methods or innovative approaches,” the feasibility of alternative methods has been assessed.

This paper presents a practical example of using multisource data collection in combination with statistical modelling, to produce statistics in a complex environment with a quite stable population. New data sources and methods require new quality framework as for example described by De Broe et al. (2021) which is highlighted in this paper.

2. How to obtain data on micro-level, with help of different sources and different methods

2.1. Introduction

The Structural statistic in EU is regulated by the IFS-regulation (EU 2018/1091). This regulation specifies, among other things, which variables must be included in surveys on the structure of agriculture in the 2020s. The survey consists of core data and eight modules, which are included in different years (see Appendix 2). The results to be delivered to Eurostat consist of microdata for each variable for every holding included in the survey or sample.

The regulation states that there are three possible sources for obtaining microdata: statistical surveys, administrative sources and other sources, methods or innovative approaches.

One objective was to reduce the number of variables collected through statistical surveys. The third method was relatively unfamiliar and the regulation and handbook provided little guidance on how to implement it.

2.1.1. Work before the survey

The work began in 2019. For each variable, discussions were held on whether a register could be used, whether the variable should be included in the survey or whether other innovative methods could be applied.

A particular challenge concerned the module on animal housing and manure management. Previous surveys had shown difficulties in this area. For example, it is challenging to design paper questionnaires that respondents interpret correctly, it is usually a high item non-response on these variables and there is a lot of existing information present in all kinds of registers, surveys and other sources. Based on these findings we decided to evaluate different methods for collection of data on the variables in this section.

2.2. Theoretical framework

There is a paradigm shift in official statistics in the sense that the traditional way of producing statistics through statistical surveys is changing. Over the past 20 years, administrative registers have been central to this shift, but alternative methods, including model-based ones, are increasingly being explored.

While alternative production methods are widely discussed, there are relatively few concrete examples of how to transition from traditional surveys to multisource approaches that are not solely based on administrative registers. This work focuses on the practical implementation of such methods without a fully established quality framework.

The evaluation of the Animal housing and manure management module began in 2019 by understanding the different variables in the sense of what is allowed

and not allowed by regulations in Sweden. Experts were consulted to better understand the practical aspects, and our ideas were continuously validated with them.

The next step was to identify the information that exists on the subject. We identified several different sources, which organisation owned the information and what we needed to obtain the data. Sources could for example be administrative registers at authorities or private companies or in some cases also the microdata from other surveys.

Based on the work done so far, we received a lot of information from different sources in 2019 and started to evaluate different methods for obtaining microdata for each variable. Even if we did not have an exact solution for each variable, we had a plan and ideas about how to tackle the problems.

2.3. Methods

2.3.1. Information from sources, regulations and other factors that affect methods

Sources included administrative registers, advisory registers, and survey data. Each source could contain information on many or only a few holdings, and a holding could appear in multiple sources. In some cases, a holding appeared multiple times within a single source (e.g. advisory registers covering multiple years).

Regulations in Sweden also influence outcomes for different holdings. Additional factors must be considered, such as climate conditions, established practices, long-term stability of certain conditions, and correlations between variables.

2.3.2. Variables in survey and methods

From the preparatory work we knew which variables pertain to practices which are not allowed due to national regulation in Sweden and which are non-existent due to for example climate conditions. For many holdings it was implicit that the results would be 0 or non-existent for certain variables as they were based on

the presence of something else in the Core part of the survey. For example, only holdings with dairy cows should have information for variables connected to animal housing for dairy cows. The work was done in a certain order based on which variables in the survey that correlated to each other.

Based on different data sources the first step was to create a large table with all data that could be relevant for the variable/variables. The table had one row and many columns for each holding. Data came from different sources, different years and could be different information from different sources.

The second step was to assess consistency across sources. If all linked sources provided the same information, that information was used. If inconsistencies occurred, sources were evaluated based on factors such as data age and relevance to the target variable. Rules were established to determine which source to prioritise. In some cases, the data directly matched the target definition; in others, it informed logical modelling.

Various methods were used to select data for each variable. Some data directly matched the target definitions, while others partially matched or showed strong correlations. Each variable required a tailored approach, considering both available data and regulatory constraints. Coverage also varied: some variables had up to 90% coverage from a single source, while others had much lower coverage.

The next step was to set values (impute) for each holding and all variables. This was done using a wide range of methods based on the specific requirements and available data for the different variables. The methods included direct imputation, imputation based on distribution, relationships and correlations, regression models between several variables and combination of different imputation methods.

Variables with consistent information from multiple sources and fewer missing values were assumed to have higher quality. Where data were sparse, quality was likely lower.

The overall quality is difficult to measure, and it differs greatly between different topics and different variables. The result must also be compared with the

alternative of collecting data through a survey (paper/web), which from experience would mean a partial non-response of up to 50% on individual variables. For some variables we can compare the result on aggregated level from the farm structure survey with other surveys for the whole country and for regions. Then we can see that the quality is good for some variables on the national level, but this does not necessarily mean that the information is correct for each holding.

2.4. Results

We have information for each holding in survey, some information of better quality and some of worse quality. In total we think the results are good enough for each holding and the data passes the validation rules that Eurostat have defined. We think that aggregated data have good quality, but it is not necessary that data for a single holding is correct.

2.5. Conclusions

2.5.1. Lessons learned

A constant problem with linking data from different sources is that the information can be linked to the holding in different ways. We learned many things from this process which will be valuable in the future. In retrospect, we might have started the preparations earlier and created the models before the survey was conducted.

It is important to clearly document which methods and models were applied to each variable. Researchers using the data must be aware that methods and data quality may vary even within a single variable.

The process of retrieving data from different organisations are not always straightforward. There could be both legal and practical or financial issues to resolve. The method entails less response burden for the holdings, but more work for the organisation conducting the survey.

2.5.2. Quality framework

This work was done in somewhat reversed order. We started with an idea and a thought that we believed in. Our tests indicated that it showed great potential. However, we did not look too much into how to describe the results of the work in terms of quality.

The framework by Daas et al. (2010) is recommended by EU (2016) for assessing the quality of administrative registers. In 2016 the framework was used for evaluating registers for FSS 2016 (Karlsson & Grönvall, 2016). However, integrating several sources need further consideration as has been shown by Waal et al. (2020).

The next step could be to look at those proposed guidelines that exist and recently have been developed. The Komuso project for example has proposed a manual called “Quality Guidelines for Multisource Statistics” (QGMSS) that could be an input. New data sources and methods require new quality framework as for example described by De Broe et al. (2021) which is highlighted in this paper.

In the system of IFS-statistics we are by regulation allowed alternative methods for creating statistics, but the quality reporting must be fitted into a fixed template which is largely based on the results being derived from statistical surveys or to some extent administrative registers.

Appendix 1: List of references

De Broe S, et al. Updating the Paradigm of Official Statistics: New Quality Criteria for Integrating New Data and Methods in Official Statistics. Statistical Journal of the IAOS. (2021) ; 37: (1). 343-360

De Waal, T., A. van Delden, and S. Scholtus. 2020. Multi-source Statistics: Basic Situations and Methods. International Statistical Review 88: 203–228. <https://doi.org/10.1111/insr.12352>

Karlsson, A., & Grönvall, A. (2016) Using administrative registers for making a sample frame for agricultural statistics-methods, techniques and experiences

Daas, P.J.H., Ossen, S.J.L., Tennekes, M. (2010) Determination of Administrative Data Quality: Recent results and new developments.

Appendix 2:

Which year core and each module in regulation 2018/1091 of the European Parliament of the council of 18 July 2018 on integrated farm statistics and repealing Regulation (EC) No 11/66 and (EU) No 1337/2011

For 2020 the core data should be collected in a census, while the modules could be sample surveys with predefined precision requirements expressed as relative standard error. For 2023 and 2026 both core and modules could be sample surveys.

Module	2020	2023	2026
Core	X	X	X
Labour force and other gainful activities	X	X	X
Rural development	X	X	X
Animal housing and manure management	X		X
Irrigation		X	
Soil mangement practices		X	
Machinery and equipment		X	
Orchard		X	
Vineyard			X