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

Ylva Olsson, Anders Grönvall, Jesper Fransson, Elin Lund

Swedish Board of Agriculture, Statistics Division, Sweden

Ylva.Olsson@jordbruksverket.se

Anders.Gronvall@jordbruksverket.se

Abstract

In the framework of agricultural statistics, a most important statistical survey is the Farm Structural Survey that has been conducted in EU since 1966 and in Sweden since 1927. The surveys of the structure of agriculture in the EU are regulated in (EU) 2018/1091 on integrated farm statistics, which, among other things, describes what data the member states should collect as well as how and when to collect it. In the regulation it is stated that the member states shall use one or more of the following sources or methods for the purpose of obtaining data on microlevel:

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

Traditionally, a statistical survey has been conducted and since 2000 Sweden has developed methods where some of the data collection is done by using administrative registers. The third option was new for this regulation and the guidelines did not give any specific help on how this could be done.
In the preparatory work for the 2020 census we evaluated the different possibilities to collect data for each variable. One part of the survey covered the field of animal stables and manure management. In total it included xx variables. From previous experiences we knew that these types of variables were difficult to collect with good quality through questionnaires in a statistical survey It is difficult to make questions regarding these issues understandable in a questionnaire, which results in a high partial non-response and adversely affects other parts of the survey and the total willingness to participate in it. For national purposes, Sweden also conducts a fertilizer and manure survey every third year which covers the national need for statistics in the field.
From the preparatory work we knew that no single administrative register would cover the entire need for information regarding these aspects. However, there is a lot of data and information available in different places, like administrative sources, information at 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 create data on micro level for each holding. The large number of variables made it a big challenge where a lot of different methods were used. The experience from this project will have a significant/great impact on the work with production of statistics in the future.

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. This entails different challenges and maybe more work for the Swedish Board of Agriculture, but less work and decreased response burden for the holding.

For a long period of time, the response rate has decreased as well as the genereal willingness to participate in surveys. This requires more follow-up telephone calls from the Swedish Board of Agriculture and it is not always the case that the farmer can answer or wants to answer.

Since 2000, the possibility to link data from registers has constantly been evaluated for each variable in the survey. This means that the definition of the variable also is evaluated so it corresponds with the target definition Because of the decreased response rate, and from 2020 the possibility to use „other sources, methods or innovative approaches", the feasibility to use different methods has also been evaluated. This paper will try to explain how Sweden for the 2020 structural survey on agriculture used a large number of sources to create a model approach for a section of the survey's variables.

This paper is showing a practical example of using multisource collection of data in combination with statistical modelling, to create 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 statistics in EU is regulated by the IFS-regulation (EU 2018/1091). This regulation describes, among other things, which variables must be included in surveys on the structure of agriculture in the 2020s. The surveys are to be carried out three times in the 2020s; 2020, 2023 and 2026. The survey consists of core data and eight modules. Variables in core data and in two modules are included all three times; the other six modules are included one or two times, see Appendix 1. 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. The final results that shall be delivered to Eurostat is microdata on each variable for each holding included in the survey/sample.

In 2020 there was a census for core data and sample survey for three modules. The regulation states that there are three different sources for obtaining data on microlevel for variables:

- statistical surveys,
- administrative sources,
- other sources, methods or innovative approaches.

There was a goal to reduce the number of variables where data were collected from statistical surveys. A statistical survey by the Swedish Board of Agriculture is done as a postal or web questionnaire combined with following up by phone. There is also extensive experience with what administrative sources can be used for statistical purposes. The third method, however, was unfamiliar ground and the regulation and handbook offered little or no advice on how to go about it.

2.1.1 Work before the survey

The work started in 2019. For each variable, there was a discussion whether a register could be used, if the variable should be included in the survey or if other innovative methods could be used. We had to decide in the winter of 2019/2020, before we comitted to the contents of the questionnaire. We know from earlier years how we could handle variables for crops, animals, labour force, other gainful activities, and rural development with help of registers and survey.

The problem was how to handle the module Animal housing and manure management. There are about 70 variables in this module. From previous surveys we knew that there are some challenges with this area. For example:

- it is difficult to include these variables in a paper questionnaire so that the respondents understand correctly,
- it is usually a high item non-response on these variables,
- in this area there are no complete administrative registers,
- 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. During the last 20-year period, administrative

registers have been the focus of the change, but alternative methods, including model-based ones, are being discussed and considered to an increasing degree.

In many contexts, alternative ways of producing statistics are advocated. There is a lot written about problems and challenges with alternative methods, but not so many concrete examples of how to go from a traditional statistical survey to something that is not administrative registers but is based on information from several sources or information channels.

This creates demands on alternative ways of producing, updating quality frameworks, etc. This work has focused on, in practical terms, using alternative methods for statistical production without having a clear quality framework to work from. The next step could be to create proposals for quality frameworks based on this practical test from reality.

2.2.1 Variables in the survey

The process of evaluating the module Animal housing and manure management started in 2019 by understanding the different variables in the sense of what is allowed and not allowed by regulations in Sweden. We also talked to experts to learn about the area and how it worked in practice and evaluated each of our thought with the experts. The section Animal housing and manure management consists of four topics that were divided into different detailed topics. Each detailed topic consists of one to twenty variables. The four topics were Animal housing, Nutrient use and manure on the farm, Manure applications techniques and Facilities for manure. Variables in each topic are presented in appendix 2.

2.2.2 Identification of sources

The next step was to identify the information that exists on the subject. We have a pretty good idea on where to find information and knowledge but needed to have more detailed information. So, we identified a number of 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. But also organisations that give advice to farmer about crops, animal, manure application.

2.2.3 Results before the survey

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, which sources were important, how many holdings we could link to each source, correlations between variables in the survey and between variables in the survey and variables in other sources.

Based on our findings we decided that all variables in the module Animal housing and manure management in some way should be calculated with help of a lot of other variables, either by imputing data from other sources combined with models or using other sources as a foundation for modelling.

2.3 Methods

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

A source can for example be an administrative register, advisers' register and data from surveys. Information from different sources was used. Each source could have information for many or few holdings and each holding could be included in one or several sources. In some sources a holding could be included many times, for example in advisers' register for different years.

There are also regulations in Sweden, both on national and regional level, that affect the results for different holdings. Furthermore, various other aspects need to be considered, such as what is possible with regards to climate conditions, practice that has been developed over the years, which conditions can remain the same for many years and correlation between different 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. In total, there were 50 variables left where we should have data for each holding in the module Animal housing and manure management. The total number of holdings in the module was 20 689.

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. The work started with Animal housing, the next step was Manure applications techniques and facilities for manure and the work ended with Nutrient use and manure on the farm. The work in all three parts was done in a similar way.

2.3.2.1 Linking data

Based on those different data sources the first step was to create a large table with all data that could be relevant for the variable/variables. Data was linked through a number of different linking variables which were updated in connection to the collecting of information for the core part of the survey. 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. Even though we have good information about linking variables, the work of linking information between different sources is a problem. These are well-known problems, and because information from administrative sources has already been linked for many years, there is a structured system to handle the various problems.

The table always includes information about number of animals and hectares of different crops from the core part of the survey. In some case some calculations have to be made, to understand if the holding has to meet requirement in the regulations.

2.3.2.2 Determining data

The second step was to see if data from different sources was consistent. For example, if all sources that was linked to the holding give the same results then we use the information.

If data from different sources were inconsistent we did evaluate the different sources by other factors like age of data, relation to target variable etc. There were a series of rules to determine which source or information to use in cases where the information from different sources was different. The values in some variables could also be input for other variables. For example, the results for variables in the animal housing topic was input in the creation of data in the manure application technique topic.

The information that was linked to the dataset could correspond exactly to the target definition on variables but it could also be information that was to be used in the logical modelling of data.

2.3.2.3 Selection of data

There were numerous different methods to select the data to use for the 50 variables for which microdata was required. In some cases, the data matched the target definition exactly; in some cases, the linked data matched some part of the target definition of the variable and in other cases there were data with high correlation to the target definition of the variable. So, for each variable what data to be used had to be identified, but also what regulations and other things that affected the possible outcome on micro level.

The linked data also provided information regarding different shares of the target population. For some variables, any source could cover up to 90% of the population, while for other variables there was considerably lower coverage.

In summary, for each variable to be imputed work had to be done to reach the target definition for the variable, but also to generate results for those holdings where there was no information in the sources from which the relevant information was linked.

We then had a table with as much information as we could get, with information for all variables for some holdings, information for some variables for some holdings and no information for some holdings.

2.3.2.4 Imputation

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
- imputation based on legal or practical conditions in real life.
- imputation based on relationships and correlations that are more or less known
- regression models between several variables
- a combination of the above imputations.

For some variables, the results were extremely good, where the data we linked could hit up to 90% of the holdings with the correct definition of the variable. And where the information did not exist it could with very good precision be imputed using other information from the holding. However, sometimes the results were less good and sometimes the quality was not possible to determine.

Holdings where we could not link any information regarding a certain variable, imputation was made using statistical imputation method. It could be done by randomly select result from similar holdings or generate results based on statistical distribution. In this imputation method, any regulations are also included as factors.

2.3.2.5 Quality

Variables where several sources were consistent and there was information to fill out many cells probably had high quality data. If there was less information from sources and many empty cells, the quality might be less good.

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. This very fact and the arrangement that micro-data must be sent to Eurostat means that the work of imputing the non-response would still have had to be done to a large extent, and probably with less available information. The main challenges identified are:

- to gain access to data and legal aspects in connection with this
- the linking of data from different sources
- the combination of different methods both in single variables and between variables
- the understanding of regulations in combination with practical situation
- the need for more knowledge on data sources
- the need for more knowledge on the area subject to survey.

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. There are numerous other problems with linking data from various sources, but that should be the topic for a separate discussion.

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. We would perhaps then have identified one or a few individual variables that we would have collected in connection with the survey that went out for other variables.

It is important to properly record which methods and models were used for which variables in the documentation. Researchers who wish to use data for all holdings in the survey must

know that the methods used and the quality of the data are not always exactly the same even for a given 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. It might take time to actually obtain the data from other organisations and there might be a monetary cost if they need time to carry out the work of retrieving the information from their systems. Maybe they request that a contract is established, determining how we may handle the data, who may work with the data and that data is confidential to everyone not working within the framework of official statistics.

In spite of significant preparations before the survey was conducted, a lot of time was required after the survey to create the models and find the best values to put in each variable. 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. However, in our case we do not only have multiple sources for collection of data, we are also combining multiple sources with statistical modelling. 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.

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 1

Which year 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

Module	2020	2023	2026
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 equpment		X	
Orchard		X	
Vineyard			X

Appendix 2

Variables in the module Animal housing and manure management

Variable		units/ categories	NE
Topic: animal housing			
	Detailed topic: bovine housing		
MAHM 001	Dairy cows	Average number	
MAHM 002	Dairy cows in tied stalls (slurry)	Places	
MAHM 003	Dairy cows in tied stalls (solid manure)	Places	
MAHM 004	Dairy cows in loose/cubicle housing (slurry)	Places	
MAHM 005	Dairy cows in loose/cubicle housing (solid manure)	Places	
MAHM 006	Dairy cows in other types of housing (slurry)	Places	x
MAHM 007	Dairy cows in other types of housing (solid manure)	Places	x
MAHM 008	Dairy cows always outdoors	Places	x
MAHM 009	Dairy cows partly outdoors (grazing)	Months	
MAHM 010	Dairy cows with access to exercise yards	Yes/no	x
MAHM 011	Other Bovine animals	Average number	
MAHM 012	Other bovine animals in tied stalls (slurry)	Places	
MAHM 013	Other bovine animals in tied stalls (solid manure)	Places	
MAHM 014	Other bovine animals in loose/cubicle housing (slurry)	Places	
MAHM 015	Other bovine animals in loose/cubicle housing (solid manure)	Places	
MAHM 016	Other bovine animals in other types of housing (slurry)	Places	x
MAHM 017	Other bovine animals in other types of housing (solid manure)	Places	x
MAHM 018	Other bovine animals always outdoors	Places	
MAHM 019	Other bovine animals partly outdoors (grazing)	Months	
MAHM 020	Other bovine animals with access to exercise yards	Yes/no	x
	Detailed topic: pig housing		
MAHM 021	Breeding sows	Average number	
MAHM 022	Breeding sows in fully slatted floor	Places	x
MAHM 023	Breeding sows in partially slatted floor	Places	
MAHM 024	Breeding sows in solid floor housing (excluding deep litter)	Places	x
MAHM 025	Breeding sows where entire surface is deep litter	Places	
MAHM 026	Breeding sows in other types of housing	Places	
MAHM 027	Breeding sows outdoors (free range)	Places	
MAHM 028	Breeding sows outdoors (free range)	Months	
MAHM 029	Other pigs	Average number	
MAHM 030	Other pigs in fully slatted floor	Places	x
MAHM 031	Other pigs in partially slatted floor	Places	
MAHM 032	Other pigs in solid floor housing (excluding deep litter)	Places	x
MAHM 033	Other pigs where entire surface is deep litter	Places	
MAHM 034	Other pigs in other types of housing	Places	
MAHM 035	Other pigs outdoors (free range)	Places	
MAHM 036	Other pigs outdoors (free range)	Months	x
	Detailed topic: laying hen housing		
MAHM 037	Laying hens	Average number	
MAHM 038	Laying hens in deep litter housing	Places	
MAHM 039	Laying hens in aviary house (without litter)	Places	x
MAHM 040	Laying hens in cages with manure belts	Places	
MAHM 041	Laying hens in cages with deep pits	Places	x

MAHM 042	Laying hens in cages with stilt house	Places	x
MAHM 043	Laying hens in other types of housing	Places	
MAHM 044	Laying hens outdoors (free range)	Places	
Topic: nutrient use and manure on the farm			
	Detailed topic: UAA fertilised		
MAHM 045	Total UAA fertilised with mineral fertilisers	ha	
MAHM 046	Total UAA fertilised with manure	ha	
	Detailed topic: manure exported from and imported to the agricultural holding		
	Net export of manure from the farm		
MAHM 047	Net export of slurry/liquid manure from the farm	m3	
MAHM 048	Net export of solid manure from the farm	tonnes	
	Detailed topic: organic and waste based fertilisers other than manure		
MAHM 049	Organic and waste-based fertilisers other than manure used on the agricultural holding	tonnes	
Topic: manure application techniques			
	Detailed topic: incorporation time per type of spread		
	Broadcast		
MAHM 050	Incorporation within 4 hours	% band[1]	
MAHM 051	Incorporation after 4 hours	% band[1]	
MAHM 052	No incorporation	% band[1]	
	Band spread		
MAHM 053	Trailing hose	% band[1]	
MAHM 054	Trailing shoe	% band[1]	
	Injection		
MAHM 055	Shallow/open-slit	% band[1]	
MAHM 056	Deep/closed-slit	% band[1]	
Topic: facilities for manure			
	Detailed topic: manure storage facilities and capacity		
MAHM 057	Manure solid storage in heaps	%	
MAHM 058	Manure stored in compost piles	%	x
MAHM 059	Manure stored in pits below animal confinement	%	x
MAHM 060	Manure stored in deep litter systems	%	
MAHM 061	Liquid manure/slurry storage without cover	%	
MAHM 062	Liquid manure/slurry storage with permeable cover	%	
MAHM 063	Liquid manure/slurry storage with impermeable cover	%	
MAHM 064	Manure stored in other facilities n.e.c.	%	
MAHM 065	Daily spread	%	x
MAHM 066	Manure stored in compost piles	Months	x
MAHM 067	Manure storage in pits below animal confinement	Months	x
MAHM 068	Manure storage in deep litter systems	Months	
MAHM 069	Liquid manure/slurry storage	Months	
MAHM 070	Manure stored in other facilities n.e.c.	Months	

1) Manure applied with specific applications technique percentage bands: (0), (>0-<25), (≥25-<50), (≥50-<75), (≥75-<100), (100).

The sum of MAHM 057-MAHM065 shall be 100 %.