# Record linkage in agricultural statistics

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

**Abstract**

The use of data from administrative registers have been used extensively in Sweden since Sweden became a part of the European Union in 1995. Integrating administrative registers with censuses and sample-surveys has been seen as a cost-effective way of producing statistics with sufficient quality.

To successfully link records from a specific unit in an administrative registers with a corresponding unit in a statistical register is therefore essential for the quality of the final statistics. In some cases, the linkage is perfect but in many cases the unit in the administrative record does not uniquely relate to the unit in the statistical register. Choices and rules taking into account the information at hand must then be used to perform the record linkage.

In this article, the outcome of using different rules for linking data from administrative registers into the statistical farm register is discussed.

**Keywords:** *Record linkage, administrative registers, multisource statistics*

## Background and aim of the paper

In this paper, we will describe how to, in a practical way, use information from different sources to improve the linkage between the population frame and administrative registers. In practice, there are several problems occurring when linking data from administrative registers. We will describe how to use extensive information from the sources in order to create better linkage to improve quality of data significantly, using the Farm Structure Survey 2016 (FSS) in Sweden as an example. FSS is in Sweden a combination of a census and sample survey where data are collected from several administrative registers as well as from several surveys.

## Creating a farm register from several sources

The Swedish Farm Register (FR) is a register of all agricultural holdings meeting predefined threshold requirements regarding the size of the agricultural area or the number of animals on the holding. The FR is updated yearly by using information from registers as well as information from different statistical surveys. When the Farm Structure Survey (FSS) is conducted, the FR is more thoroughly updated. The FSS is conducted every third or fourth year and is regulated by EU law. The latest FFS was done in 2016 and the content in the survey, for example, includes characteristics of land, crops, livestock, organic farming, irrigation, labour force, other gainful activities, support for rural development, soil and manure management practices. The aim of the survey is to give decision-makers statistical information on agriculture to plan, monitor and evaluate the impact and efficiency of current and possible policy changes. Such changes can, for example, relate to the environmental impact of agriculture and to sustainable farming methods.

In Sweden, the FSS has consistently been conducted as a census for some characteristics and as a sample survey for some characteristics. This means that all holdings receive a questionnaire but that it is more extensive for a sub-sample of the population. Characteristics that correspond to the creation of thresholds are collected from all holdings while information on other characteristics, for example labour force, are only collected from a sub-sample.

##  *Creating sample frame for FSS*

The frame for FSS is created using the FR for the previous year, updating it with information from registers as close to the execution of the actual survey as possible before the actual survey. The registers used for creating the sample frame are:

* IACS (Integrated Administrative and Control System), which contains information about the application for agricultural subsidies farmers can make. In IACS there is information on crop areas for all individuals who applied for area based agricultural subsidies. For each agricultural holding there could be several persons applying for subsidies. In some rare cases there could also be several holders connected to one person applying for subsidies. The majority of holdings with agricultural land applies for agricultural support.
* CDB (register for cattle), which contains information about each head of cattle, where it is located and who is responsible for the animal. An agricultural holding can have several registered locations for cattle and in some cases cattle at one location could be connected to several agricultural holdings.
* Animal registers, which contain information about locations with different animals (sheep, pigs, poultry). There is no information about the actual number of animals at the location, but there is information about the maximum number of animals at the location. An agricultural holding can have several locations with animals and in some cases animals at one location could be connected to several agricultural holdings.
* Business register, which contains information about all business entities. This register is used to detect horticultural holdings, which are not covered by IACS or animal registers.

When creating the final frame for FSS, the FR for the previous year is the base. By using deterministic linkage between the FR for previous year and newest version of registers we create a preliminary FR for 2016 which also is used as a sample frame for the FSS. For register objects that could not be linked to the FR, a new holding is created. The deterministic linkage method (or exact linking) assumes that information has a perfect match.

 **Figure 1. *Creation of sample frame for FSS2016***



##  *Creation of holdings in FSS 2016 and FR 2016*

The FSS in Sweden is a mixed survey using information from different sources. All holdings receive a questionnaire with questions regarding information needed for linking variables such as personal identification number, production place number and customer number for agricultural subsidies. In the questionnaire, there are also questions regarding total agricultural land and the number of animals (sheep, pigs and poultry). A sample of holdings receive a more detailed questionnaire with questions about irrigation, labour force, other gainful activities, support for rural development, soil and manure management practices etc. The regulation of FSS stated that we only had to do a sample survey, which is the reason for making a sample. Holdings not included in the sample still received the questionnaire to improve the accuracy of some estimates as well as to retain the possibility of dividing estimates to lower levels. One important reason is also to update the FR.

The sample is divided into two different categories, one for legal entities and one for private holdings. This was done in order to allow for questions about labour force to be adapted to the legal status of the holdings. For 2016, Statistics Sweden also made a sample survey with questions including management practices. Due to this, we divided the two sample categories into two more groups depending on their presence in the survey conducted by Statistics Sweden. The holdings which were included in the survey conducted by Statistics Sweden did not receive questions answered by that survey. We ended up with four different groups for which we made a separate questionnaire. A fifth questionnaire was sent out to those who were not included in the sample. This questionnaire was less extensive.

The questionnaires were sent out in the end of May with reference day for the number of animals and the crop area set to the first Thursday in June. The collection period was June-November. In August, registers were updated, with the latest information about the status in the beginning of June.

When information from all sources is correct and straightforward, the linking of data from registers to the population is unproblematic. However, in approximately 8-10 % of the holdings, there is information that does not match or that is illogical. It could be that information in the registers is old and not updated or it could be the farmers have misunderstood the questionnaire. It could also be that the situation on the holding is more complex than what could be fit into the survey. To make the best possible link using the information gathered from all different sources, we use all the information to link in the best possible way.

Figure 2. Creation of holdings in FSS2016 and Farm Register 2016



For all holdings, there is information about the main holder and information about the co-holders. The co-holder is a person who is connected to the holding, for example a wife, a son or daughter who works or is active on the holding. It could also be an employee. The person answering the questionnaire can mark other persons as linked to the holding. We also ask for information used in linking variables such as customer number for subsidies or production place numbers for animals.

Therefore, in theory, we can link data from registers to the holding based on either the main holder or the co-holder. We can also link through customer number or production place numbers.

*2.2. Creation of variables for FSS 2016*

When the linking of objects from different sources has been completed, the information on characteristics is collected from all the different sources and put into a complete dataset.

Figure 3. Creation of variables for FSS2016



## Problems in practice - linking data from different sources

The problems with linking should be quite small if everything was as expected. When all linking variables to all different registers are included in the questionnaire, there should not be many problems. However, in practice there are several problems occurring due to different causes. For example:

* The person applying for subsidies for a given piece of land is not the person responsible for the holding that uses the land. In most cases it is a question of owning the land and renting the land.
* The respondent claims in the questionnaire to have ceased with agricultural activities, but still has animals and/or areas in the registers.
* The information on the questionnaire intended for linking results in a failed or impossible link.
* The same information intended for linking exists on multiple questionnaires.
* The respondent claims to have sold the holding to another person but the other person does not exist in registers.
* The respondent does not provide any information which could be used to establish links, even though such information exists.

The problems expressed above can be divided into four different categories:

1. The farmer has misunderstood the questionnaire or the aim of the survey. The holders do not see themselves as farmers in the statistical definition. They only do it as a small part time activity but they reach the statistical threshold for being a farmer.
2. The register is not updated with the latest information about holdings or there is a mismatch between the person that put the information in the administrative register and the person actually responsible for the farming.
3. The information in the registers or in the questionnaire is incorrect.
4. The situation on the holding is more complex than can be described by a questionnaire.

Of course, there is also a mix of the different problems above. Therefore, we have a problem with the information the respondents have put on the questionnaire and that it does not correspond with the information in the registers.

In total, we found 8-10 % out of 70 890 holdings in the 2016 frame that had conflicting information (that provided conflicting information./where the information in the questionnaire conflicted with the information in registers.).

If we look only to the holdings that indicated that they are no longer agricultural holdings on the questionnaire, there were 13 643 of them. Out of these holdings approximately 5 000 could immediately be classified as a non-agricultural holding as they had no connection to any register anymore. The remaining 8 500 holdings had information which was in some way confusing or conflicting when data from registers were compared with data from the questionnaire.

The conclusion is that pure deterministic linking does not work

## Using multiple variables from different sources to create better linkage

*4.1. Different linking methods*

In theory we can read about mainly two different linking methods:

1. Deterministic linkage (exact linking) – This method uses exact matching between different registers. For example personal identification numbers give the exact match between two registers.
2. Probabilistic linkage – This method uses information on a greater number of matching variables and is used when unique identifiers are not available, or are deemed unreliable.

There is of course a set of alternative linkage methods that are used in practice. The result from this study is a development of the deterministic method. This due to the fact that the actual linkage based purely on unique identifiers does not take care of some basic problems in the datasets.

*4.2. Practical linking in Farm Structure Survey 2016*

Because of the specific problems arising in the FSS, the use of only deterministic linkage would not be sufficient to get a good view of the statistical population. However, as a first step we use a deterministic linking approach where unique identifiers are present in the population and in registers. The identifiers in the population are put on the questionnaire to enable the respondent to make alterations to them. However, there is information on the questionnaire that can be used to improve linkage when the identifiers are not updated etc. or the information in the register is not in correct for the actual purpose of the survey. To reduce bias in linking we can use other information put on the questionnaire to improve the quality of the linking procedure. This is what we call “The black box of linking” in Figure 2.

We know that using deterministic (exact) linking based only on the identifiers linking the variables would give us an overestimate of the number of agricultural holdings. This is mainly due to the fact that there is a time gap between when registers are updated and when the survey is conducted. For example, the process of discontinuing an agricultural enterprise is often not completed over night. It is done in a successive way where the actual starting point is when one holding rents the land from another holding. The old holder continues to apply for subsidies for some period, which means that if we do not consider this we would have two holdings using deterministic (exact) linking. The main problem is not due to missing links but to what extent we should use the exact linking or use other information the probabilistic method is not an option. Our By using all auxiliary information given in the questionnaire and also in other registers we can reduce the risk of incorrect linking.

Figure 4. Example of linking by using auxiliary information



The way we handled this was to divide those variables that could indicate a problem into certain results on a nominal scale. For example, holdings has the value “yes” if their answers on the questionnaire indicate that they have closed down and “no” if there are no such indications. In total, we had five such variables.

1. Holdings closed down (yes, no) according questionnaire
2. No new owner assigned on questionnaire (yes to main holder on other holding, yes to co-holder on same holding, yes to co-holder on other holding, yes to holder that does not exist in FR, no)
3. Link from IACS to holding is possible (yes to new main holder, yes to previous main holder, yes to both new and old main holder, yes to co-holder, no)
4. Link from CDB to holding is possible (yes to new main holder, yes to previous main holder, yes to both new and old main holder, yes to co-holder, no)
5. The respondent has answered the questionnaire (yes, no).

By creating a hierarchical structure, we could divide the holdings into one of 150 groups depending on how they have answered the questionnaire and what information we could receive from registers.

Figure 5. Creating a hierarchical structure of information

By this method, we identified a group of holdings that had conflicting information from different sources. For example, there was a group of 2 300 holdings which were indicated in the questionnaire as not engaged in agricultural activities any more. The holding was sold, rented or in other ways no longer in the possession of the main holder. Therefore, they had not answered the rest of the questionnaire. They did not give us any information regarding whom they had transferred the holding to. For those 2300 holdings, however, we had information from registers about a second holder on the holding. Therefore, we could conclude that for the vast majority of these holdings they had transferred the holdings to a person that we already had a connection to on the same holding. Our solution was then to change the main holder on the holding and to categorize these holdings as non-respondents. For the census variables, we imputed those values which were not found in a register. In practice this means that the respondents at those holdings did not understand the questionnaire correctly, which means that we should think about changing it for the next survey.

Questionnair is completed

Links to CDB exists

Links to IACS exists

Holdings closed down according questionnarie

So on group level (grouped by information from questionnaries and data from registers) we could evaluate the conflicting information given from different sources and in most cases we could make a good solution on how to handle the holdings ending up in better results in survey/census but most of all a better farm register form future surveys.

## References

Wallgren, A, Wallgren, B (2014), Register-based Statistics, Second edition, Wiley

Sakshaug, JW, Antoni M. (2017), Total Survey Error in Pracice, Wiley