USING ADMINISTRATIVE REGISTERS FOR MAKING A SAMPLE FRAME FOR AGRICULTURAL STATISTICS - METHODOLOGIES, TECHNIQUES AND EXPERIENCES

Ann-Marie Karlsson, Anders Grönvall
Swedish Board of Agriculture, Statistics Division
551 82 Jönköping, Sweden
Ann-Marie.karlsson@jordbruksverket.se
Anders.Gronvall@jordbruksverket.se

DOI: 10.1481/icasVII.2016.f36c

ABSTRACT

Swedish official statistics in the agricultural area has since 1995 been based upon the extensive use of administrative data. However, in most cases it is not advisable to produce statistics direct from administrative registers. This paper reviews the methodologies and techniques used in order to ensure sufficient statistical quality when using administrative sources together with the Farm Structure Survey for the creation of the statistical farm register. The Farm register (FR) is used as a sampling frame for all agricultural statistics based on holdings.

In the paper we will show how the frame can be updated regarding the holdings constituting the frame as well as regarding the characteristics for example area of different crops, rented land and number of animals needed to stratify samples. The results show that integrating administrative registers with surveys is a cost-effective way of updating the frame while reducing the burden on respondents. Various aspects of the quality dimensions “accuracy”, “comparability” and “coherence” are the key issues for development in order to improve the quality when integrating registers and surveys.However if extensive, systematic work is integrating registers and surveys, the quality in these dimensions can be high. It is shown that there is at least as much need for work with improving quality, methodological studies and quality assurance for statistics based on administrative registers as for statistics based on sample surveys. When using administrative registers, the integration phase where data from several sources are integrated into a new statistical register is central for improving quality.

Keywords: register-based statistics, administrative registers,

1. Background and aim

At least from 1965 and onwards agricultural statistics in Sweden has been dependent on an updated Farm register (FR). The FR consists of all agricultural holdings in Sweden as well as variables needed for creating a typology consisting of type of farm, size of farmand region. The variables includedifferent kinds of animals and crops, rented land and some general information about type of holding and the holder. The typology is used to stratify effectively and thus minimise the sample in almost all agricultural sample surveys. Up until 1996the FR was updated by a yearly census where farmers were asked about all crops and animals.

When Sweden joined the EU in 1995 administrative registers of all farms applying for EU-subsidies where made available for statistical purposes. Subsequently several studies were made to see how the registers best could be used for agricultural statistics.Selander et al. (1998) and Wallgren& Wallgren (1999) for example compared IACS (the register for area-based subsidies) from 1996 with the objects in the FR from 1995. The results showed that 88 per cent of the objects matched, 4.9 per cent of the objects had multiple links between the registers and 6.9 per cent could not be found in IACS. Of the objects in the IACS register on the other hand, 2.9 per cent could not be found in the FR.As a result of the studies work the following years were focused on integrating the registers and FR. Several questionnaires were sent to subgroups of farmers asking them about keys in IACS.

In recent years Dias et al.(2016) has shown,with the example of Portugal, how alternative methodologies to a traditional population-census could be evaluated. They conclude that since there are advantages and disadvantages to all methods, it is important to make a systematic evaluation so that the trade-offs between options could be taken into account in the decision making process. It is stressed that registers might not have all the content needed; instead the registers may need to be combined with traditional surveys. The findings in the studies of Wallgren & Wallgren (1999) also showed that it is not possible to produce the FR directly from administrative registers. However they showed that administrative registers could be used for updating the objects in the FR as well as give information on some variables, for example crops.

As a result,starting in the year 2000 the FR was no longer updated with an annual full census, instead itwasupdated with an integrated use of registers and surveys in the years for which EU require the member states to conduct a Farm structure survey (FSS) according to regulation (EC) 1166/2008.In the years in-between, the FR is updated mainly from registers, together with information from a sample survey on animals. In recent years additional registers have been used,for example administrative registers on cattle, sheep, poultry and pigs.

When using an administrative register there is at least as much need for methodological studies and quality assurance as for statistics based on statistical surveys. However, the quality deficiencies and the approaches to investigate and resolve them differ.Several studies, for example Wahlgren & Wahlgren (2014),Laitila et al. (2011), Daas et al. (2009, 2010) and Agafitei et al.(2015) have discussed quality frameworks for using administrative registers for statistical purposes. This includes the quality of the register itself, the possibilities of integrating administrative registers into statistical registers and how to document the quality of the statistics produced. There is also a discussion about which part of the process to focus on. Holmberg (2015) summarises that there is an ongoing theoretical development on how to assess the quality of administrative data, but that more work is needed for example regarding linkage errors and cowerage errors.

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 (SJV, 2016)

1.1 Aim of the paper

This paper will describe how the FR in Sweden is updated using administrative registers in combination with some census data. The quality achieved by using theadministrative registers in combination with a statistical survey will be compared towhat quality could be achieved by updating FR using only a census and using only administrative registers.The framework recommended by Eurostat (2016) presented by Daas et al. (2010) will be used to highlight the quality in the administrative registers.

The quality will be described according to the quality criterias stipulated in regulation (EC) No 223/2009 on European statistics. Furthermore aspects of cost-effectiveness, privacy and response burden will be addressed.

2. Updating the Farm register (FR)

Figure 1 showsthat the overall method for creating the FR is to make a maximal sampling frame (2) withcombining the objects in the FR from the latest FSS year (1a)by all new possible objects that can be found in registers (1b). Next a questionnaireis sent out to the maximal sampling frame(3), and a current FR is made (5)by combining the information in the maximal sampling frame with information from the questionnaire(4). In the end the variables needed for thetypologyare added.

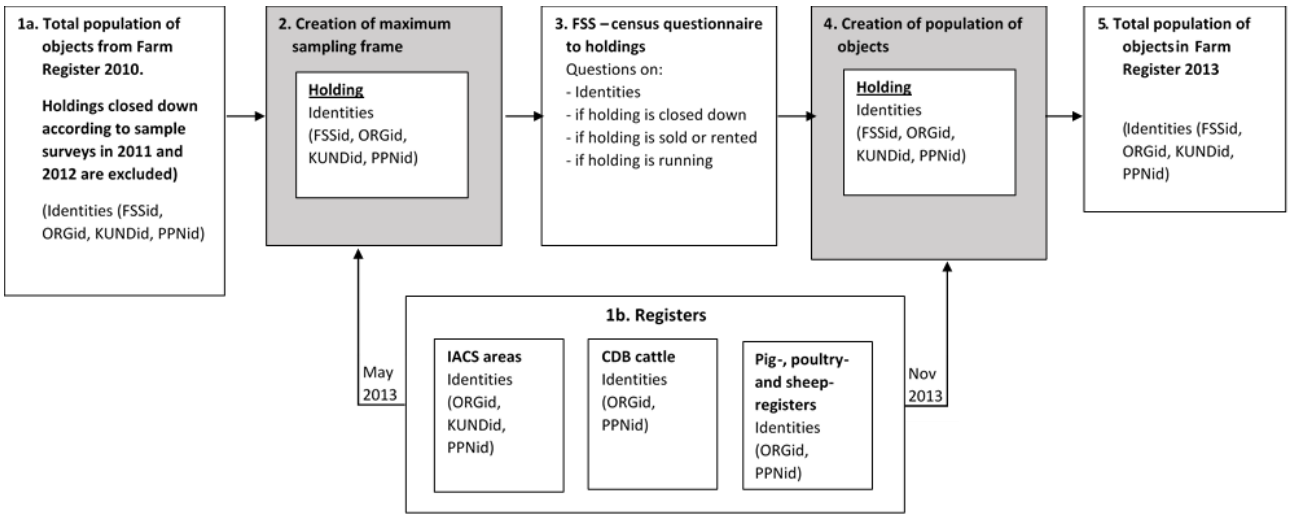

Figure 1:*Creation of the Farm register 2013.*

The target population of the farm register is the same as the target population of FSS. The variables of animals and crops in FR are also defined in the same way as the variables for animals and crops in FSS. I.e. producing the FR means doing a large part of the work needed for FSS.

2.1 Creating the maximal sampling frame

In the beginning of 2013 the total population was taken from the 2010 Farm Register.The indicators related to the source as described in the hyper-dimensions of (Daas et al., 2010) are used to describe the quality of the registers. The Swedish Board of Agriculture is the supplier of the data sources for all registers except the Business register (BR), where the supplier is Statistics Sweden. The Board is also the responsible NSO for agricultural statistics, so privacy and security rules are in place. Historically the registers have always been delivered in time. All persons and organisations in the registers as well as in the FR have a unique organisational number (ORGid). The ORGid is therefore always a linkable auxiliary key between registers.

The IACS register is highly relevant. It consists of all holdings applying for area based subsidies in May each year. In IACS about 60 codes for crops are included. It has unique keys (KUNDid) and the linkability is high. It has clear definitions and the information is well checked. However there is an under-coverage since about 10 per cent of the holdings do not apply for or are eligible for area based subsidies. Following the definition in FR a holding might contain several KUNDid. IACS could be used both for updating the register and for a large proportion of the variables.

The registers for poultry, pigs and sheepas well as the cattle database are aimed at tracing animals in case of an outbreak of a contagious animal's disease. Each production-place with these animals have a unique key (PPNid). The production-place is related to a person or legal entity with an organisational number (ORGid). One holding can relate to several production-places. The information in the registers differs. The poultry-and pig- registers mainly holds information on the number of animals that could be held at the production-place.

In the sheep-register additional information on the number of sheep in December are added by a questionnaire. There is a unit non-response of about 24 per cent. This register also has an over-coverage of about 20 per cent. It can be assumed that the over-coverage of the poultry- and pig-registersis the same. The under-coverage is small since the PPNid needs to be reported when animals are slaughtered. These registers could be used to update the objects in the maximal sampling frame but the quality is not sufficient for the variable values.

In the cattle-databaseeach head of cattle has an identifier thatis related to a production place (PPNid). There is, however, no distinction in the register between dairy-cows and cows for meat-production. Information from a second auxiliary register of milk deliveries is used to obtain the required information. It is assumed that if the holding delivers milk in the month of the reference day, the cows on the holding are all dairy cows. This approach will resultin a small over-estimation of dairy cows and subsequently an underestimation of cows for meat-production. However a sample survey conducted in 2002 showed that the error was less than 1per cent. The cattle database is therefore used both for updating the objects and for the variables needed for cattle.

The use of registers means that thepostal questionnaire sent to the farmers did nothave to include questions on crops and cattle since this information is available from registers. Only questions on the number of horses, poultry, pigs-, and sheep were included.Since there could be multiple to multiple objects in the different registers, the order of merging is decided by specific rules. Each holding in FSS has a set of related keys (for example Orgid and KUNDid). The keys are ordered by their quality. If there are multiple possibilities of linking, the key ordered the highest i.e. with the best quality is used first.

The linkages made in the spring 2013 showed that 64 036 objects could be linked to the registers while11 689 holdings in the FR 2010 could not be found in registers. 3 004 of them had not been found in registers 2010 either and 8 685 are holdings that existed in registers 2010 but not 2013. Information from sample surveys done in 2011 and 2012 indicate that 800 holdings had closed down during the period. 868 additional holdings were found in the cattle register, IACS or the pig-, poultry-, and sheep registers. 934horticultural enterprises in the business register were added to find horticultural holdings that might fit into the frame. In the end the maximal sampling frame amounted to 77 527 holdings. By creating the sampling frame this way there will be an over-coverage in the frame. The under-coverage is assumed to be small as long as the holdings were present in 2010.

When the questionnaire was created it included questions regarding whatkeys (ORGid, KUNDid, PPNid) that were related to the holding and whether it was sold, rented out, closed down or still running. If the holding was sold or rented out it was askedwho was now running the farm. It is assumed that most of the 8 685 holdings that could no longer be found in the registers are no

longer farmers, but some of them are likely to be the same as the 868 holdings that were found in registers but not in the register from 2010. This together with the linkage process could mean that the same holding might get more than one questionnaire. Whether this has occurred is therefore a question in the questionnaire. The questionnaire also includes questions on animals and rented land, variables that could not be found in the registers. The variables grazing land and arable land are included as a reference to the area of rented land.

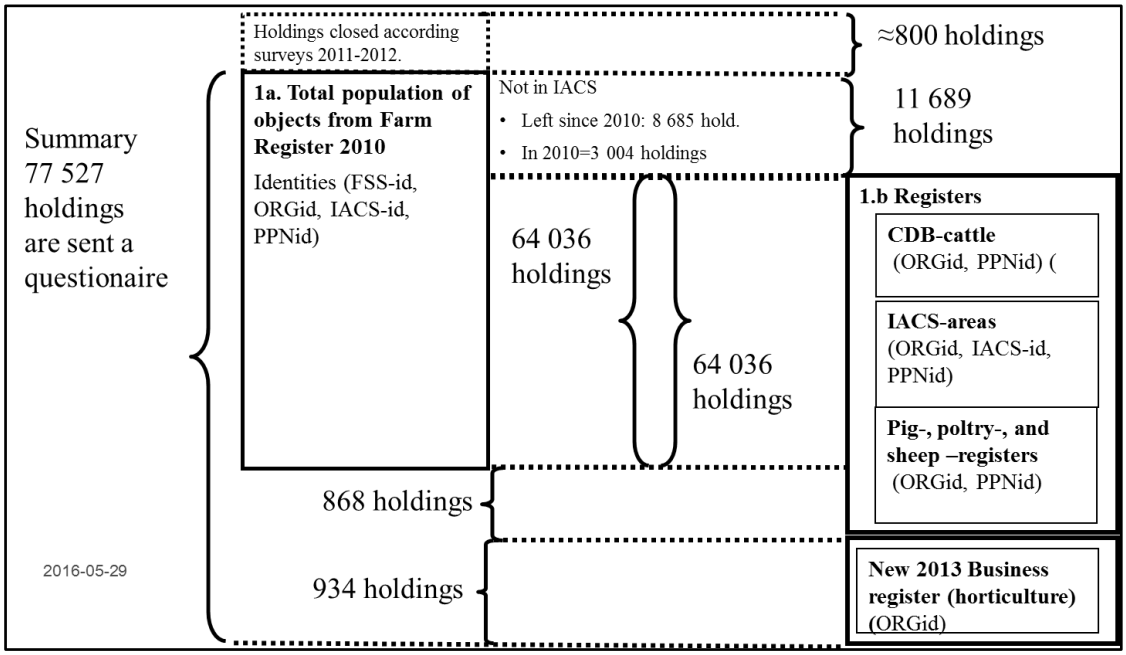

Figure 2.*Creation of the maximal sampling frame.*

2.2 Results from 2013

In November when the answers from the questionnaire were processed, updated data was gathered from the registers and related to the reference time. The information from the questionnaire and the registers wasmatched and the population of objects for 2013was created. In the questionnaire the set of keys and the quality of the keys related to the holding were updated. The linkage was then made using the same principles as when creating the sampling frame. The linkage resulted in 67 126 holdings in the FR for 2013.

613 duplicates were found. The duplicates were mostly holdings in FR 2010 that were not found in registers included in the group "left since 2010". The duplicates were due to linkage errors.

The holdings in the group "left since 2010" were mainly growing crops. Out of the 8 684 holdings 3 125 had been closed down and 3 626 had been taken over by another holder. I.e. only 16 per cent were still farmers so there is a large over-coverage in this group. Out of the 3 004 holdings in the group "In 2010" 989 holdings had closed down and 541 had sold the farm to someone else. I.e. 58 per cent were still running the holding. It could be concluded that those holdings that for several years are running their business without applying for subsidies continue to do so.

The results regarding the horticultural register showed that 30 per centwere still running their holdings.574 holdings had closed down and 77 had sold it to someone else. For the holdings in this group it is more common to close down the business than letting someone else continue it. Out of the holdings in the pig-, poultry-, and sheep groups about half were still running their holdings and out of the large group that matched with IACS 98 per centwere still running their holding.

When the population was created, the variables were connected to the objects using the established links of keys. The information on areas and crops given in IACS as well as information about cattle given in the cattle register are considered to have high quality since the information is subject to extensive controls. For example, the holding is requested to give information about the total area of arable land in hectares as well as the total area of rented arable land in hectares. If the total area of arable land differs from the summarised areas of arable land used for different crops in IACS, the IACS figure will be used. As a result, the answers in the questionnaire wereonly used to calculate the share of rented arable land at the holding. The divergent answers could also be a sign that the creation of objects had not been fully successful. Consequently the information was used to improve the creation of objects.

3. Comparing quality between registers, census and multiple source

Every 10:th year a full census needs to be made meeting the target-population of FSS according to regulation (EC) 1166/2008. The census includes the variables in the FR. The alternatives which are interesting to compare are different methods for updating the FR in the two times during the 10 year period when the FSS is stipulated to be made as a sample survey. The three alternatives compared are a:

1. combination of census and registers i.e. the current method described in section 2,
2. register based approach, using registers in combination with the FSS-sample survey,
3. full census based method using a census and not any of the registers available.

3.1 Cost-effectiveness, privacy and response burden

Privacy and the principle that information given for statistical purposes should not be used for administrative purposes are fundamental when producing statistics (Eurostat 2016). On the other hand, administrative registers may be used for producing statistics. In Sweden, the Official Statistics Act (2001:99) addresses data disclosure and the Secrecy Act (1980:100) addresses the confidentiality of individual information. In the case of the FR, individual data from administrative registers are protected by the Secrecy Act (1980:100) when used for statistical purposes, regardless of whether the data would be public or not from the body responsible for the administrative register.

Using administrative registers is also a way of reducing the burden on respondents. At our user-meetings the Union of Swedish Farmers states that it is favourable to share information between governmental bodies in order to make it possible for farmers to only provide the same information once. The hours spent filling out the present set of questionnaires are calculated to 5 800 hours. If no administrative registers where used the questionnaire would also need to have information on all animals and all crops and the total number of hours spent could be calculated to 9 600 hours. If only administrative registers were used the time would be 0. However compared both to the current method used and the full census the quality of the register would be lower. The samples for surveys in the years 2014-2020 would need to be larger. We have calculated those extra hours to 9 500 hours. This includes larger samples for the sample surveys of crops, rented land, fertilisers and the census 2020.Over the 10 year period the difference between register-based statistics and the current method regarding response-burden is low.

The total cost for the current method used in 2013 was 750 000 euro. The cost for a full census through postal questionnaires could be estimated to 1 100 000euro and for producing the statistics solely from registers to 450 000 euro. However there would be an additional cost for the larger samples in the years 2014-2020. Producing statistics from administrative registers is cost-effective in relation to postal questionnaires. From a cost-effectiveness point of view it can be seen that a large proportion of the cost is due to handling paper questionnaires. In 2013 only 20per cent

answered using the web service. I.e. the most obvious way of saving money would be to persuade farmers to answer through the web.

3.2. Quality criteria

Relevance refers to the degree to which statistics meetcurrent and potential needs of the users. The updated FR is harmonised with the target population and some of the variables in FSS. I.e. the needs from EU-legislations are met. The FR also can be used to disseminate statistics for small regions or groups as well as combining data in new ways without restrictions. It could be assumed that a full census would meet the same quality criteria. If the FR were updated only by registers itwould still fulfil the needs from EU-legislation regarding FSS. However since animals would be sampled it could only be used for producing statistics on NUTS3 sometimes NUTS2 level.

Accuracy refers to the closeness of estimates to theunknown true values.For register-based surveys, key issues are integration errors as well as how well the definition of objects and variables in the registers correspond with the required definitions in the statistics. The problems of matching different sources should not be underestimated. There is a risk for over-coverage in the FR. Parts of the same holding or the same areas could be counted several times. It is possible that a landowner, who is no longer cultivating his land and who belongs to the 8 685 holdings that could not be found in the register, answers the questionnaire on the basis of the land that he owns without cultivation. At the same time, the tenant of the same land who has applied for subsidies also states that he uses the land. Using only a census would lead to an under-coverage of the FR since the holdings not applying for subsides might be lost and new holdings more difficult to find. Using only registers for creating objects might also lead to an under-coverage since holdings not in the subsidy systems would be excluded.

For *timeliness* i.e. the period between the availability ofthe information and the event or phenomenon it describes there are no differences between approaches.Wallgren & Wallgren (2014) stress that timeliness is often a problem when using administrative registers. However in the case of IACS and the cattle-database this is not a problem. IACS data are available in May the reference year and the cattle-database is updated constantly.

Punctuality which means if release dates are kept,*accessibility and clarity* that has to do with the user's possibilities to access and compare data and *coherence* i.e. if the information is put together and presented in a logical way present no differences between alternatives.

Concerning *comparability*, or the possibility to compare results over time, problems might occur. Statistics based on administrative registers are dependent on changes that the statistical bodies may not be able to predict. For example, the change in the CAP in 2005 meant that the holders applying for subsidies in 2005 would be eligible for subsidies in subsequent years. The changes in the CAP affected the statistics in several ways, for example the number of holdings increased. Two explanations to this increase could be found. Firstly it was thought that small farms that had not been eligible for subsidies now applied for subsidies and were consequently incorporated into the population of holdings. The change in administrative rules thus improved the quality of the FR, correcting a previous under-coverage. Secondly, it could be assumed that some landowners applied for subsidies although the land in practice was cultivated by a neighbouringholder, leading therefore to over-coverage in comparison with the definitions in the FR.This problem would be the same for the register based approach but would not occur for the alternative based on full census.

The quality in terms of comparing results from registers and surveys are high because of the wide-ranging work done to merge different registers. In this sense the quality would be lower in the two other alternatives.

1. Conclusion

To conclude the use of registers is cost-effective and reduces the response burden for the holdings. Regarding the quality criteria of relevance, there are advantages to integrating registers with surveys and censuses when collecting data for the FSS. On one hand, in solely register-based surveys, the register may not cover the requested variables. On the other hand, in solely statistical surveys, the questionnaire would be expensive and cause a high response burden.Regarding the dimension of timeliness, the registers used for the FSS are updated and available earlier than results from a survey or census would be. When comparing the accuracy of the results, there are several aspects to consider.However, integrating registers and surveys as well as only using a statistical survey is considered to give accurate results. The coherence and comparability is assumed to be the highest when registers and surveys are integrated because of the wide-ranging work done to merge different registers. The availability is good in all three alternatives.

The paper shows that different aspects of accuracy are the key issues to consider in order to improve the quality when integrating registers and surveys in the FSS.

REFERENCES

Agafitei, M., Gras F., Kloek W, Reis F., Vaju S. Measuring output quality for multisource statistics in official statistics: Some directions. *Statistical Journal of the IAOS, vol. 31,* 2015. *no 2, pp. 203-211.*http://dx.doi.org/10.3233/sji-150902

Dias C., Wallgren A., Wallgren B., Coelho P. Census Model Transition: Contributions to its Implementation in Portugal.*Journal of Official Statistics. Volume 32, Issue 1, Pages 93–112,* ISSN (Online) 2001-7367, DOI: 10.1515/jos-2016-0004.

Daas P, Ossen S, Vis-Visschers R, Arends-Toth J (2009). *Checklist for the Quality Evaluationof Administrative Data Sources.* Statistics Netherlands Discussion Paper, 09042.

Daas, P.J.H., Ossen, S.J.L., Tennekes, M. (2010) Determination of Administrative Data Quality: Recent results and new developments. *Proceedings of Q2010 European Conference on Quality in Official Statistics*, Statistics Finland and Eurostat, Helsinki, Finland.

Eurostat (2016) *Quality Assurance Framework of the European Statistical System Version 1.1.*http://ec.europa.eu/eurostat/documents/64157/4392716/qaf_2012-en.pdf/8bcff303-68da-43d9-aa7d-325a5bf7fb42 Read:2016-05-22.

Holmberg A. (2015) Discussion. *Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 515–525,* http://dx.doi.org/10.1515/JOS-2015-0031

Laitila T Wallgren A Wallgren BQuality Assessment of Administrative Data Research and Development – *Methodology reports from Statistics Sweden 2011:2*Örebro: Statistics Sweden.

Selander, R., Svensson J., Wallgren, A., Wallgren, B. (1998) *How should we use IACS data? Administrative registers in an efficient statistical system.*Environmental and regional statistics, Statistical Report September 1998Örebro: Statistics Sweden.

SJV (2016) *Evaluation of the quality of registers used for FSS.* Unpublished. Jönköping. SJV.

Wallgren, A & Wallgren, B (1999). *How can we use multiple administrative sources?: administrative registers in an efficient statistical system - new possibilities for agricultural statistics? : statistical report October 1999.*Örebro: Statistiska centralbyrån.

Wallgren, A. & Wallgren, B. (2014*). Register-based statistics: statistical methods for administrative data*. 2.ed. Chichester: John Wiley & Sons.