

Using administrative registers for agricultural statistics - methodologies, techniques and experiences

Karlsson, Ann-Marie
Swedish Board of Agriculture, Statistics Division
Vallgatan 1
551 82 Jönköping, Sweden
E-mail: ann-marie.karlsson@sjv.se

Widén, Marie-Louise
Statistics Sweden, Agriculture
Klostergatan 23
701 89 Örebro, Sweden
E-mail: marielouise.widen@scb.se

Abstract

Swedish official statistics have long been based upon the extensive use of administrative data. However, in most cases it is not advisable to produce statistics direct from administrative registers since these registers are not adapted to the needs of the statistics. This paper reviews the methodologies and techniques applied in order to ensure sufficient statistical quality when using administrative sources for the creation of farm structural statistics i.e. the Farm Structure Survey (FSS). The results show that integrating administrative registers with censuses and sample surveys is a cost-effective way of producing statistics while reducing the burden on respondents.

Regarding the quality dimension “content” integrating registers with surveys and censuses presents some advantages when collecting data for the FSS. Regarding “timeliness”, the administrative registers used for the FSS are updated and available earlier than results from a statistical survey.

This paper shows that various aspects of the quality dimension “accuracy” are the key issues for further development in order to improve the quality when integrating registers and surveys in the FSS. The quality discussed in terms of “coherence” and “comparability” is high when registers and surveys are integrated because of the extensive work done to merge different registers.

It is shown that there is at least as much need for quality work, 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.

1. Introduction

Swedish official statistics have long been based upon the extensive use of administrative data. In official agricultural statistics, administrative data are used for collecting data on farm structure, farm incomes, farm economics and animal production. Administrative data are also used to update the Swedish Farm Register. In the Farm Register, all objects and the information variables are gathered for the population of Swedish farm holdings. The administrative sources used mostly consist of information given by farmers for the purposes of receiving EU subsidies, bookkeeping, animal health and taxation.

In most cases, it is not advisable to produce statistics directly from administrative registers since these registers are not adapted to the needs of the statistics. Overall quality, object sets, object definitions and variables need to be checked and, thereafter, it may be necessary to do a certain amount of processing so that the register fulfils the quality requirements of the statistics. There is at least as much need for quality work, methodological studies and quality assurance for statistics based on administrative registers as for statistics based on statistical surveys. However, the quality deficiencies and the approaches to investigate and resolve them differ.

In the article, the concepts “sample survey” and “census” will be used for surveys where the data collection is carried out directly from the objects. The concept “register-based survey” will be used when the microdata are collected from administrative registers.

This paper reviews the methodologies and techniques applied to ensure sufficient statistical quality when using administrative sources from the following dimensions:

1. The possibility of using administrative registers from a cost-effective, legal, organizational and technical point-of-view as well as aspects of privacy and the response burden on farmers;
2. The quality components of the statistics.

The Farm Structure Survey (FSS), where a combination of registers and questionnaires are used, will be used as a case study to discuss the different aspects.

2. Integrating registers with a census and sample survey in the FSS 2007

When Sweden became a member of the European Union in 1995, several studies were carried out to investigate the consequences of using administrative registers created for administrating the Common Agricultural Policy (CAP) for statistical purposes (Selander et. al 1997; Selander et. al 1998). For example, the objects in IACS (the register for area-based subsidies) from 1996 were compared with the objects in the Farm Register from 1995. The Farm Register was, at the time, based on yearly censuses. The results showed that 88 percent of the objects matched and that 4.9 percent of the objects had a multiple linkage between the registers. Of the objects in the Farm Register, 6.9 percent could not be found in IACS. On the other hand, 2.9 percent of objects in the IACS register could not be found in the Farm Register.

As a result of the studies, the annual census was replaced in 2000 by an integrated use of registers, a census and a sample survey in the years for which the EU require

the member states to conduct a FSS, i.e. 2003, 2005, 2007, and 2010. In the years in-between, the Farm Register is created solely from registers, except for a sample survey for animals. The creation of objects i.e. holdings in the FSS for 2007 is shown in Figure 1.

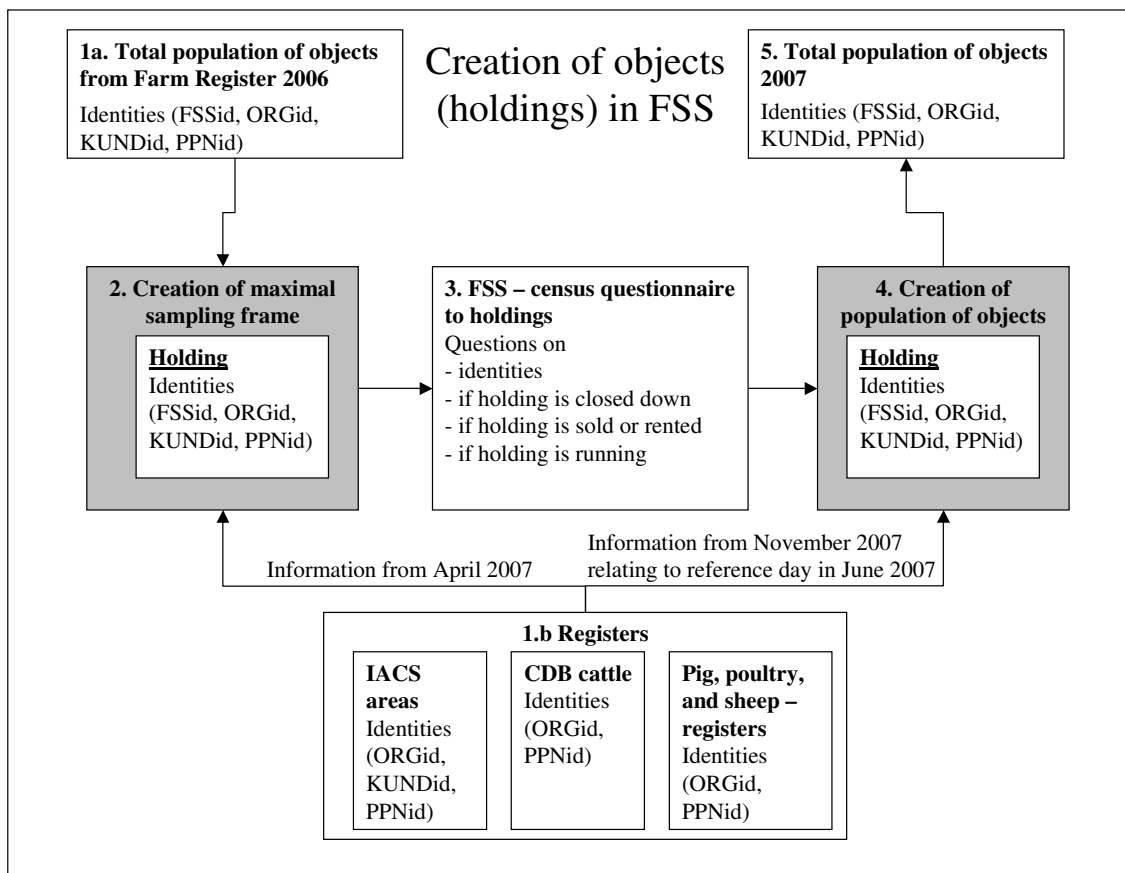


Figure 1. Creation of objects in FSS

In April 2007, the total population is taken from the 2006 Farm Register. Several identities are linked to each holding:

- FSSid that defines the object i.e. the holding in the FSS
- ORGid that defines the legal unit. There can be multiple ORGid for one FSSid
- KUNDid that defines the applicant identity in IACS. There can be multiple KUNDid for one FSSid
- PPNid that defines the cattle holder in the central cattle database and other animal-related registers. There can be multiple PPNid for one FSSid.

In order to create the maximal sampling frame, ORGid, KUNDid and PPNid were used for merging the total population in the 2006 Farm Register with the IACS register, CDB cattle register and pig, poultry and sheep registers. Since there could be multiple to multiple objects in the different registers, the order of merging is decided by specific rules. For example, if there is diverging information about what identities from a register constitute one holding, the identities are used in the order of ORGid, KUNDid and PPNid.

In 2007 about 80 000 holdings were created and constituted the maximal

sampling frame;

69 500 holdings were a match between the 2006 Farm Register and the different administrative registers.

- 8 500 holdings in the 2006 Farm Register had not applied for subsidies or existed in any register in 2007.
- 2 000 holdings could be found in the registers for 2007 but could not be linked to a holding in the 2006 Farm Register.

During the summer of 2007, a questionnaire was sent out to all holdings in the maximal sampling frame. In relation to the creation of objects in the questionnaire, the holders were asked what identifiers (ORGid, KUNDid, PPNid) they had in different registers and if they were still farmers. If they were not, they were asked to whom they had sold or rented their holding or if the holding was no longer maintained. It is assumed that most of the 8 500 holdings that could no longer be found in the registers are no longer farmers. However, this is checked by the questionnaire.

In November when the answers from the questionnaire are processed new updated data will be gathered from the registers and related to the reference time. The information from the questionnaire and the registers will be merged and the population of objects for 2007 will be created. Holdings are linked by set of specific rules. Links from the administrative registers, for example links between PPNid and KUNDid given in IACS, have a higher priority than links between the PPNid and KUNDid given in the questionnaires. The administrative registers are thus considered to have a higher quality than the questionnaire since the identities are used for the payment of subsidies to each farmer. In specific cases, however, the individual answers in the questionnaires are used to create holdings. In order to get qualitative time series, as many of the 2 000 new holdings as possible are connected to the old 8 500 that could no longer be found in the administrative registers for 2007. The rules of linking are constantly reviewed in order to create a population that fulfils the requirements for the FSS.

When the population is created, the variables are connected to the objects using the established links of identities. The information on areas and crops given in IACS as well as information about cattle given in the CDB cattle register are considered to have high quality since the information is subject to extensive controls. The creation of variables is shown in Figure 2. As can be seen in the figure, the FSS is constructed as a census for some variables and as a sample survey for some other variables (specially required by the EU). If there are diverging answers between the questionnaire and the registers, the information in the registers is considered to be of higher quality. 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 summarized areas of arable land used for different crops in IACS, the IACS figure will be used. As a result, the answers in the questionnaire will only be 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 has not been fully successful. Consequently the information is used to improve the creation of objects.

When creating the variables, some variables are processed. IACS has for example more details on crops than are needed for the FSS. As a result, several crops in IACS are added to one crop in the FSS. The registers of sheep, pigs and poultry could not be used for creating variables since information about the number of animals required for the FSS is not included in the registers. In the case of the sheep register, the number of sheep is included but the information could not be linked to the reference time in the FSS.

The FSS has been conducted in 2003, 2005 and 2007 and will be conducted again 2010. In the years in-between, the objects of the Farm Register are created solely from registers i.e. steps 2 and 3 in Figure 1 are excluded. The variables are almost completely taken from registers except for a sample surveys for animals.

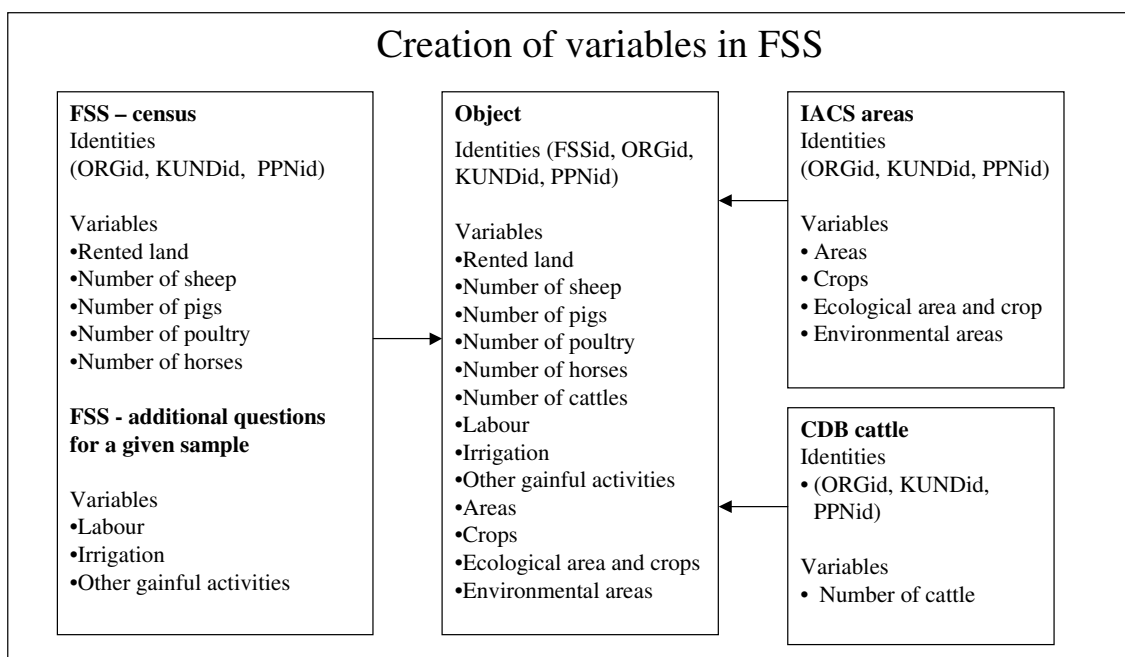


Figure 2. Creation of variables in FSS 2007

3. Aspects of privacy, the response burden, costs and technical issues

Privacy and the principle that information given for statistical purposes should not be used for administrative purposes are fundamental when producing statistics (Eurostat 2007, UN 2007). 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 FSS, individual data from administrative register are protected by the Secrecy Act (1980:100) when used for statistical purposes, regardless of whether the data would be public or not if requested by the body responsible for the administrative register. Legally, the possibilities of using registers are addressed in the Official Statistics Act (2001:99).

Using administrative registers is also a way of reducing the burden on respondents. Administrative sources should be used in order to avoid duplicate requests for information (Eurostat, 2007). The Union of Swedish Farmers have also stated that it is

favourable to share information between governmental bodies in order to make it possible for farmers only to provide the same information once (LRF, 2006).

Producing statistics from administrative registers is cost-effective in relation to postal questionnaires or interviews.

In order to use administrative data, organisational issues and technical issues must be addressed. There is a close cooperation between the 25 statistical authorities that are responsible for different areas of statistics in Sweden. The Swedish Board of Agriculture (the Government's expert authority in the field of agricultural and food policy,) and Statistics Sweden also cooperate to address problems in agricultural statistics.

4. Aspects of quality in register statistics

Different aspects of the use of administrative registers for statistical purposes are comprehensively addressed by Wallgren and Wallgren (2007). Wallgren and Wallgren state that there is at least as much need for quality work, methodological studies and quality assurance for statistics based on administrative registers as for statistics based on sample surveys. However, the quality deficiencies and the approaches to investigate and resolve them differ.

In general, the quality of statistics relates to all aspects of statistics which are relevant to how well they meet the users' needs for statistical information. The users' needs must always be the reference point. Regarding quality, the FSS should first of all fulfil the requirements of the European Union (regulation 571/88). However, the FSS is also used for national needs. The FSS is used as a sampling frame for most of the agricultural surveys. It is also a part of the business register and is used for analyses and research. The following discussion on how to address the quality aspects of statistics based on administrative registers will follow the quality aspects as outlined by Statistics Sweden, which are contents, accuracy, comparability and coherence, timeliness and finally availability and clarity.

Since there are usually budget and time constraints, the users' needs must always be in focus when decisions are made on how to prioritize between different dimensions of quality. The result of the prioritization process will assist the decisions regarding what collection methods should be used. Wallgren and Wallgren (2007) state that relevance errors and integration errors are the main quality issues in register-based surveys. In the case of the FSS, it is considered that integration errors are the most important ones to be addressed while relevance errors are of minor importance. In the cases where data in the registers are not relevant or non-existent, they are replaced or completed by a census or sample survey.

Firstly, it must be understood, as is pointed out by Wallgren and Wallgren (2007), that the information in an administrative register is collected for a specific administrative purpose. IACS has the purpose of fulfilling the requirements for CAP and the CDB cattle register is made to monitor animal healthcare. Consequently the quality of the registers and the way data are processed at the administrative bodies will depend on administrative requirements. Information that is of low value for the administrative purpose will in general be of lower quality. On the other hand, surveys

based on samples or censuses also have quality flows. Some respondents might not understand the question, respond carelessly or not at all (Wallgren and Wallgren, 2007). Quality aspects of a sample survey or census are related to sampling errors, non-response and measurement errors. The problem of creating a sampling frame that reflects the target population is a problem for register-based statistics in the same way as for censuses and sample surveys.

When using administrative registers, the integration phase where data from several sources are integrated into a new statistical register is central. Accordingly, in register-based statistics, there are integration errors. This might include coverage errors, matching errors, missing values due to mismatch and aggregation errors. In order to monitor and improve the quality, it is important to have good knowledge of the registers (Wallgren and Wallgren, 2007). When integrating register-based surveys with censuses and sample surveys, all kinds of errors might occur.

4.1 Contents

The contents of a register relate to what variables are collected. Comparing the users' need for data with what is included in registers is a way of indicating the quality. For the FSS, the users require specific information and the information not included in administrative registers has to be collected by censuses or sample surveys. As can be seen from Figure 2, a large proportion of the variables exist in administrative registers. Administrative registers might thus be considered to have high quality in terms of content.

4.2. Accuracy

Accuracy describes whether the statistics correctly estimate or describe what it was meant to measure and if data are reliable. For a sample survey, accuracy is indicated by how well the sampling frame corresponds with the target population, sampling and measurement errors, and by focusing on how well the definition of variables is understood by the respondents. 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 definitions of objects and variables in the administrative registers are a result of the administrative requirements. For the FSS, this means that the definitions of objects do not have a one-to-one relationship between the Farm Register and the administrative registers. There might be objects belonging to the FSS population that do not exist in any of the administrative registers. As a result, matching errors might occur and, as previously stated, the phase of integrating objects is a critical.

Wallgren and Wallgren (2007) advocate the use and comparisons of several sources in order to improve the quality of register-based statistics. This is done in the FSS, as can be seen in Figure 1 and 2.

However, the problems of merging different sources should not be underestimated. The approach used in the FSS could lead to over-coverage. There is a risk that 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 8 500 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.

The definition of variables is also a presumed problem. However, for the FSS, the definitions in the administrative registers and the definitions in FSS are in most cases similar. Usually the variables in the administrative registers are more detailed than what is required for the FSS so variables from registers have to be added or processed in order to get the demanded variables in the FSS. One example is the number of dairy cows, where the register gives information about birthdate, sex and type for each type of cattle at a specific place of production on a certain date. Processing of information from a third register on milk deliveries is also necessary to obtain the required information. It is important to know the quality and definitions of the register in order to achieve the required result. For the holdings not matching with IACS, the register variables are imputed.

The quality of the variables in the administrative registers is considered sufficient for the measurements required, compared to asking the holders directly. The holders are motivated to give exact measurements of the areas they cultivate or the number of animals they possess since they may otherwise lose their subsidies. This means, for example, that the accuracy of the area of wheat is most likely much better than if the farmer simply estimated the area without any sanctions.

4.3 Coherence and comparability

Coherence refers to whether the information is put together and presented in a logical way. A specific piece of information should be comparable between times, regions or size groups. Furthermore, good quality in this dimension indicates that it should be possible to merge the information and use it in combination with other registers. Common definitions, classifications and methods are important measures to achieve coherence. In register-based surveys, it is usually difficult to influence the definitions, classifications and methods but they need to be documented and followed so that the impact for the statistics on change can be monitored.

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 possibilities of receiving subsidies for areas of grazing lands increased. The change also meant that the holder 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 that 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 FSS, correcting a previous under-coverage. It was, for example, shown that the number of horses and sheep as well as holdings with horses and sheep often held on small holdings increased.

On the other hand, it could also be assumed that some landowners applied for subsidies although the land in practice was cultivated by a neighboring holding, leading therefore to over-coverage in comparison with the definitions in the FSS.

Wallgren and Wallgren (2007) indicate that the way of solving this problem is in general the same as in statistical surveys, that is to try to link time series, to estimate the effect of the changes and explain this for the users. The quality discussed in terms of comparing results from registers and surveys are high because of the wide-ranging work done to merge different registers.

4.4 Timeliness

Timeliness is often indicated by the length of time between the reference period and the availability of the results. Wallgren and Wallgren (2007) stress that timeliness is often a problem when using administrative registers.

In the case of agricultural structural statistics, however, this is not the case. A preliminary version of the IACS register showing the applications for area-based subsidies is available as early as in April of the reference year. Consequently, information in the registers for the reference year can be used to update the maximal sampling frame. Throughout the year the register is updated and, in order to be able to incorporate the updated information when creating the population, information is also gathered from the registers in November.

The CDB cattle database is updated with information for a specific reference day a couple of weeks later. As a result, the quality aspects of availability and timeliness are higher for the administrative information than for a census or sample survey with their own data collection.

4.5 Availability

Availability describes how accessible the statistics are for the users. The statistics can be made accessible in many ways, for example, in the statistical database, publications, etc. In the case of the FSS, the quality dimension of availability is the same regardless of whether the data are collected from registers or from the objects directly. The results from the FSS are sent to Eurostat as microdata and the results are published in Sweden as well as in the EU.

5. Conclusions

To conclude this paper, the method of producing the FSS by integrating registers with a statistical survey will be compared to the exclusive use of registers and also to the exclusive use of a statistical survey. The results are summarized in Figure 3. From this, it can be seen that the use of registers is cost-effective and reduces the response burden for the holdings.

Regarding the quality dimension of content, there are advantages with integrating registers with surveys and censuses when collecting data for the FSS. On the 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 so extensive that the respondents may not answer all questions.

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.

Quality dimension	Exclusively register-based	Both registers and surveys	Exclusively statistical survey
Cost-effective	+++++	+++	+
No response burden	+++++	+++	+
Content	++	++++	++
Timeliness	+++++	++++	++
Accessibility	+++	++++	++++
Accuracy	+++	+++++	+++
Coherence/comparability	+++	++++	+++
Availability	++++	++++	++++

Figure 3 Quality aspects of the FSS

When comparing the accuracy of the results, there are several aspects to consider. However it is considered that integrating registers and surveys as well as only using a statistical survey should 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 further development in order to improve the quality when integrating registers and surveys in the FSS.

References

- Eurostat (2007), European Statistics Code of Practice, September 5 2007
http://epp.eurostat.ec.europa.eu/portal/page?_pageid=2273,1,2273_47141302&_dad=portal&_schema=PORTAL
- LRF (2006). Varför krångla till det? Rapport av LRF juni 2006.
- Selander, R., Svensson, J. (1997) Metodrapport 1997:1, Statistiska Centralbyrån 1997. IAKS-utredningen.
- Selander, R., Svensson J., Wallgren, A., Wallgren, B. (1998) How should we use IACS data? Administrative registers in an efficient statistical system - new possibilities for agricultural statistics. Environmental and regional statistics, Statistical Report September 1998 Statistics Sweden.
- Statistics Sweden (2001). Meddelanden i Samordningsfrågor 2001:1. Quality definition and recommendations for quality declarations of official statistics.
- UN 2007, Principles governing international statistical activities, September 5: 2007.
- Wallgren, A., Wallgren, B. (1999) How we can use multiple administrative sources. Administrative registers in an efficient statistical system - new possibilities for agricultural statistics. Environmental and regional statistics, Statistical Report

October 1999 Statistics Sweden.
Wallgren, A., Wallgren, B. (2007) Register-based Statistics. Administrative Data for
Statistical Purposes.