**The usefulness of quality frameworks when deciding on replacing surveys with administrative registers**

Anders Grönvall, Ann-Marie Karlsson

Swedish Board of Agriculture, Sweden

In this paper the framework published by Statistics Netherlands (Daas et al.) and the framework published by Statistics Sweden (Laitila et al.) are discussed in the context of the decision-making process of whether to use an administrative register on sheep instead of a survey to produce sheep statistics.

From the perspective of our case it was found that the frameworks are similar. Daas et al. has some advantages concerning evaluating the legal base of using the register and the register itself. Laitila et al. has some advantages concerning evaluating production process quality. The greatest usefulness of the systematic approach was in evaluating the administrative registers themselves. Further development is suggested for indicators evaluating the administrative register in relation to the goal, for the quality in the long run and for exit-plans if the register is discontinued.

**1. Background and aim of the paper**

It is a difficult and important decision whether or not to replace a statistical survey with data from administrative registers. This is especially the case for official statistics where comparability is an important component of quality. Therefore, when Sweden became a member of the European Union in 1995, and administrative registers in the agricultural area were created, studies were carried out to investigate the consequences of using administrative registers for statistical purposes [1]. The studies showed that integrating administrative registers with censuses and sample-surveys could be cost-effective ways of producing statistics with sufficient quality. The integration phase where data from several sources were integrated into a new statistical register was seen as essential for achieving sufficient quality. As a result of the studies, Swedish official statistics in the agricultural area has from then on been based upon extensive use of administrative data.

In recent years several studies have discussed quality frameworks for using administrative registers for statistical purposes. A conclusion is that there has been a theoretical improvement on how to assess the quality of administrative data [2]. At the same time it has been pointed out that compared to the maturity of statistical research regarding the methods of for example sample surveys, research on methods regarding administrative registers has just started [3].

To use or not to use an administrative register is a strategic decision. A decision is a choice between different alternatives in order to achieve a goal. A decision can be seen as a process and thus divided into several steps [4]. The steps summarised in this paper are problem-definition, i.e. defining the difference between the situation today and a desired situation, setting up and evaluating different alternatives, choosing between alternatives, implementing the choice and evaluating the result.

The aim of this paper is to discuss the usefulness of the concept developed by Laitila et al. [5] and Daas et al. [6] for the strategic decision on whether to count the number of sheep in December with a sample survey or an administrative register. Suggestions for areas where further theoretical development might be needed will be made.

**2. Quality frameworks in relation to deciding on how to measure the number of sheep**

On dimension level the concepts of Laitila et al. and Daas et al. are to a large extent overlapping. However, there are some differences in the way dimensions are systemised and how quality is assessed. Daas et al. sees three dimensions: the administrative register in itself, metadata about the source and the data in the source. Indicators are developed within each dimension. Laitila et al. discusses quality in three dimensions: output data quality, input data quality and production process quality. They also stress that the purpose of the register is important for what dimensions are to be used. Is 1) the quality of the register good enough to be used directly for producing statistics; 2) can it be used after processing and 3) can it be used to only to improve the productions process. In the third case the dimension of production process quality is used. They also group the dimensions related to the work process, that is information from the administrative authority, data editing of the source, integrating the source with the statistical register and improvement of the production process.

*2.1 What is the goal? What is the problem?*

The quality framework is discussed within the context of the process of a making a unique strategic decision on whether to base the statistics on the number of sheep in Sweden in December on administrative registers or on a survey.

The statistics regarding sheep is used for the Swedish statistical Farm register (FR) and is regulated by Swedish legislation. It is stipulated that every third year FR should be updated through a census in June. The years in between, the number of sheep is estimated by a survey in June. EU requires counts for the number of sheep in June and December. The national and EU-requirements[[1]](#footnote-1) can be summarised as that the number of sheep should be counted in December and June. EU stipulates that the sheep should be divided into the categories ewes, lambs and rams and the sampling errors for the results if a survey is used should not exceed 2 % (with a confidence interval of 68 %) of the total number of sheep. Other sources than surveys are allowed if the statistics will have at least equal quality to that of the survey.

The sheep are now surveyed by a sample survey of 7 000 farms in June. The current estimation of the number of sheep in December is of insufficient quality and needs to be replaced. To make a survey of 7 000 farms in December would induce a total cost for respondents calculated to 30 000 euro. Furthermore there is an increasing demand from Swedish farmers and the government to reduce the burden on respondents.

The goal is therefore to find the best way of estimating the number of sheep in December. The quality should be in line with EU-demands, i.e. of the same quality as the survey made in June.

*2.2. Setting up and evaluating alternatives*

The alternatives to evaluate are to make a new survey in December or use information from one or several administrative or statistical registers to reach the goal. Found registers concerning the number of sheep were the administrative sheep register, the administrative slaughtering register, registers from the certifying bodies regarding organic production of sheep and membership registers in breeding associations. The slaughtering register, which can give information on the number of slaughtered sheep, is regarded as an auxiliary register to help modeling and is evaluated as such.

The indicators related to the source as described in the hyper-dimensions of source and metadata of Daas et.al. and of information from the administrative authority of Laitila et.al. is shown in figure 1. The dimensions from the figures are referred to in by the indicator number in the text. The Swedish Board of Agriculture is the supplier (S1) of the data sources of the sheep register and the slaughtering register. The board is also the responsible NSO for agricultural statistics. The suppliers of the other registers are the bodies for organic certification and the breeding associations. About 20 % of the sheep are held with organic production methods and the number of holders in the breeding associations does not cover the total production. (S2, A1, A7) These two registers will therefore not be evaluated further in the text, but are summarised in figure 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sheep register | Slaughtering register auxiliary source | Register from bodies for organic certification | Registers from breeding associations |
| Evaluation of registers according to the of hyper-dimensions of source and metadata *Daas et al. (2010)* | | | | |
| Source dimensions |  |  |  |  |
| S1. Supplier | +++++ | +++++ | +++ | ++ |
| S2. Relevance | ++++ | ++ | + | + |
| S3. Privacy & security | +++++ | +++++ | ++++ | + |
| S4. Delivery | +++++ | +++++ | + | ++ |
| S5. Procedures | +++++ | +++++ | +++ | + |
|  |  |  |  |  |
| Metadata dimensions |  |  |  |  |
| M1. Clarity | +++++ | +++++ | ++ | + |
| M2. Comparability | ++++ | +++ | ++ | + |
| M3. Unique keys | ++++ | +++ | + | + |
| M4. Data treatment | +++++ | +++++ | ++ | + |
|  | | | | |
| *Information from the Administrative authority (Indicators of output and input data quality) according to Laitila et al.* | | | | |
| *Relevance* |  |  |  |  |
| A.1 Relevance of population | ++++ | ++ | + | + |
| A.2 Relevance of units | ++++ | ++ | + | + |
| A.3 Relevant keys | ++++ | +++ | + | + |
| A.4 Relevance of reference time | ++++ | ++++++ | ++ | ++ |
| A.5 Study domains | ++++ | ++ | + | + |
| A.6 Comprehensiveness | ++++ | +++++ | ++ | ++ |
| A.7 Updates | +++ | +++++ | +++ | +++ |
| A.8 Delivery time | +++++ | +++++ | ++++ | ++++ |
| A.10 Punctuality | ++++ | +++++ | +++ | ++ |
| A.11 Comparability of time | +++ | +++++ | ++++ | ++ |
| *Accuracy* | | | | |
| B1.Primary key | +++++ | +++ | . | . |
| B2 Foreign key | +++++ | +++++ | . | . |
| B3 Duplicates in the source | ++++ | +++ | . | . |
| B4 Missing values | +++ | +++++ | . | . |
| B5.Wrong values | +++ | +++++ | . | . |

Figure 1. Evaluation of registers according to the hyper-dimensions of Daas et al. and Laitila et al.

Regarding the dimension of relevance (S2) in December every year the animal holder should count the number of sheep on his or her production place and report the result to the sheep register (A5, A8). There is no distinction between ewes, rams and lambs in the register (A5). The number of sheep in June is not reported (A8). The slaughtering register is a register based on the slaughter houses reporting the number of animals that have been slaughtered each day and from what production place the animals has been transported (A1-A10). This register does not take into account the number of animals slaughtered on the farm.

Privacy and security issues are regulated by law. (S3) The aim of the sheep register is to provide a possibility to trace animals from one production place to another in case of an outbreak of a contagious animal disease. The production place (a building for animals mainly staying indoors, or a field if the animals are outside) is assigned a number. Each animal keeper could have several production place numbers (PPN) but the number could never be moved from the location.

The basis for the sheep register is regulation (EC) no 21/2004. The base for the slaughtering register is (EC) 2006:815 and SJVFS 1998:127. In Swedish regulation it is stipulated that administrative registers might be used for official statistics. SFS (2001:99). Confidentiality is ensured by regulation SFS (2009:400). (A1)

Regarding the dimension of delivery (S4) there is no cost to use the data source, except the few hours spent extracting the information. There are routines for transferring data between the statistics division and other divisions at the Swedish Board of Agriculture. The delivery can be made on demand and a request might take 1-2 weeks to expedite. The sheep register data are available in March-April for the December count. The data is delivered in Excel format or in text format. Delivery of data from the slaughtering register follows the same routines and data are available a couple of months after the slaughtering has taken place. (A9, A10)

Regarding procedures (S5) there are at the Swedish Board of Agriculture routines for sharing information about registers. If data is not delivered on short notice either the statistics will be delayed or the figures will be estimated based on the latest known figure.

Regarding the dimension of clarity (M1) the population of the register is all holdings that have sheep at any point during the year. Those holdings should report the number of sheep in December. The one variable reported is clear and simple, the number of sheep at the production place. The comparability with the definitions in the statistics is good for the number of sheep. However when national and EU-demands are summarised the sheep need to be divided into lambs, rams and ewes. (A4, A5)

Comparability is important. (M2) When comparing the sheep register with the statistical census or the survey, there is a large difference between the number of sheep in December and the number of sheep in June. Most lambs are born in the beginning of the year and slaughtered in the second half of the year, so the total number of sheep is far less in December than in June. The sheep register takes into account all animal holdings regardless of the size of the farm, while a threshold exists in FR and other statistics that excludes the smallest farms. The slaughtering register offers possibilities for bridging the gap between the number of sheep in June and December by deducting those slaughtered. On the individual level, however, this presents the problem that sheep can be sold between holders during the period before they are slaughtered. I.e. subtracting the slaughtered sheep of one holder from their number of sheep in June will not necessarily yield the correct number of sheep of the holder in December. Home slaughtering and sheep that for example die as a result of predators or accidents are also not included. (A1, A5, A7)

Even though one holding in FR can have several production places, one production place cannot have several holdings at a specific moment in time. There are unique keys, identifiers (M3, B2, B1) between the slaughtering register, the sheep register and FR.

Regarding data treatment the data (M4) is checked by the Board. However, if the sheep register is not correct there are no sanctions for the farmers. This indicates that the quality is lower than for those registers where the farmers can be sanctioned financially. There are routines at the Board for sharing information about changes in the register. We cannot in detail analyse duplicates in the source, missing values or wrong values until we compare the register with our own statistical register in the next step. (B3-B5)

The third hyper-dimension introduced by Daas et al. is the data dimension that focuses on the quality aspect of the data in the source or as put by Laitila et al. indicators of input data and production process quality. The indicators of Laitila et al. are still to a large extent on register level but for the indicators of Daas et al. we need to a higher extent to base our analyses on our approach of using the registers together with modeling to achieve the statistics. I.e. the quality is assessed on the full approach, not only the registers themselves. In the frameworks compared there are 10 dimensions in each that are used to evaluate the quality of the data. (Figure 2)

Information from the sheep register and the slaughtering register is linked to FR. The linking of the registers is made through different steps. Data from the slaughtering register is first linked to the sheep register using the holding numbers (PPN) as identifier. The link between the slaughtering register and the sheep register is good since both registers has PPN as their primary key. Data from the slaughtering register is in addition to the PPN also the number of slaughtered adult sheep and lambs each month. The data in the sheep register is data on the total number of sheep at a certain date in December. The combined register is then linked to FR using both PPN, and an organisational ID. Both identifiers are continuously updated in FR for linkage purposes.

There is some missing data from the register and some data that could not be used since the farmer has counted the sheep on other dates than December. There are also a small number of holdings that could not be linked. It is therefore not possible to use the number of sheep directly from the register. Instead a model using straight expansion in different strata is applied. [7] It entails adjusting for missing values in the sheep register based on the number of sheep in June at a first stage. Then we use the number of sheep in June and the slaughtering statistics as auxiliary information to calibrate the weighting factors.

In practice we also want to distinguish between the number of ewes and other sheep in December, which cannot be achieved using merely data from the sheep register where all animals are counted as a single category. Therefore we use the information from the slaughtering register and estimated home slaughtering to estimate these figures.

Regarding technical checks there are readability and metadata compliance as quality indicators. The data in the sheep register is of very good technical quality and all data can be accessed (Q1). The metadata definition is the total number of sheep in December. However, in practice the number reported in some cases concerns other dates.(B4) Regarding relevance of variables they are relevant and can be compared with other variables (A3). The register could also be used to improve the frame of the June survey even if it is not good enough for a December estimate.

|  |  |  |  |
| --- | --- | --- | --- |
| Daas et al. | Model of the number of sheep | Laitila et.al. | The  Sheep  register |
| Q1. Technical checks | +++ | C1. Undercoverage in source | +++++ |
| Q2. Over-coverage | +++ | C2. Undercoverage in farm register | +++++ |
| Q3. Under-coverage | ++++ | C3. Overcoverage in source | +++ |
| Q4. Likability | +++ | C4. Over coverage in farm register | +++++ |
| Q5. Unit non-response | +++ |  |  |
|  |  |  |  |
| Q6. Item non-response | +++ |  |  |
| Q7. Measurement | ++++ |  |  |
| Q8. Processing | ++++ |  |  |
| Q9. Precision | ++++ |  |  |
| Q10.Sensitivity | ++++ |  |  |

Figure 2. Quality of the data for the statistical purpose

Regarding coverage (Q2, Q3, C1- C4) there is an over-coverage of about 20 %, which means that 20 % of the holdings are too small to be included in the farm register and counted as farmers. In a few cases there could be holding numbers that are not in use anymore. As the register is very well updated and there is a strict law stipulating the animal holders’ obligation to report and our knowledge by comparing with the number of sheep in the farm census there is no under-coverage assumed. All holdings that have sheep are registered in the sheep register.

Regarding linkability (Q4) about 7.5 % of the holding numbers from the sheep register are not possible to link to the farm register. 1.2 % have data on the actual variable (number of sheep in December). The remaining 6.3 % have either not reported to the register at the latest animal count or have reported the number of animals on a date after 31 of December. Out of the total number of holdings there is a link of 72.5 %. This corresponds to 90.5 % of the holdings when the over-coverage is removed.

There is a unit non-response. (Q5, Q6, B4) As the data from the register actually only consists of two variables (the number of sheep and the date of counting them) that are closely linked, the unit non-response and the item non-response give the same answer. 23.8 % of the holdings have missing or incorrect data (e.g. data not corresponding to December). Out of these, 17.5 % was linkable data while 7.5 % was not possible to link. Since there is only one item, the item and unit non-response is the same.

Regarding measurement, when making the calculation we used an expansion method to create weights for each responding unit in the frame. At the first step we made a straight expansion using only the number of missing data from the register. In the next step we also used auxiliary information from the FR and slaughtering register to calibrate the weights.

In processing of data we do not make any adjustments or imputations of data from registers. However, there may be some data from registers that are linked to an incorrect holding in the FR.

Regarding precision the standard error of the estimate of the number of sheep in December is about 2.4%. Regarding sensitivity there is about 9.5 % of empty or incorrect cells in the data. Furthermore, there are 7.5 % of the data from the sheep register that do not link to the FR. The figures are summarised in Table 1. ( Q9)

*2.3. Choosing between alternatives*

The choice should then be made whether the sheep register can be considered a reliable source and if the model gives the same quality as the survey did and if the figures from December could replace the existing model.

The results from the model are ambiguous. The sheep register as well as the slaughtering register are registers that presumably will remain of the same or better quality for the foreseeable future. In 2011 the quality of the model is not good enough to replace the survey. In 2012 it is.

**Table 1: Evaluation of registers**

|  |  |
| --- | --- |
| Number of sheep in June 2010 (census): | 564 922 |
| Number of sheep in June 2010 (register): | 567 618 (before calibration) |
| Number of sheep in December 2010 (register): | 377 024 |
|  |  |
| Number of sheep in June 2011 (sample survey): | 622 711 (MSE=1, 6%) |
| Number of sheep in June 2011 (register): | 600 347 |
| Number of sheep in December 2011 (register): | 397 846 |
|  |  |
| Number of sheep in June 2012 (sample survey): | 610 534 (MSE=0, 9%) |
| Number of sheep in June 2012 (register): | 613 067 |
| Number of sheep in December 2012 (register): | 411 892 |
|  |  |
| Number of sheep in June 2013 (census): | 576 769 |
| Number of sheep in June 2013 (register): | 578 687 (before calibration) |
| Number of sheep in December 2013 (register): | 372 474 |

*2.4. Implementing the decisions and evaluating the results*

In relation to the goal it can be concluded that the register could be used to report the number of sheep in December in 2013, although the quality needs to be checked constantly.

**3. Discussion**

In relation to the decision making process the quality frameworks are naturally mostly used in the phase of choosing between alternatives. The first phase, defining the problem and clarifying the goal, should not be forgotten. Quality in a broader sense includes requirements of user needs, institutional preconditions at the organisation producing the statistics as well as the production process itself. Furthermore, evaluation of options might differ if the main goal is to reduce response burden as opposed to if the main goal is to reduce costs. It is important to have a phase in the decision making process that is open-minded and also explores alternatives. It is important to have a clear understanding of what the goal is when evaluating the alternatives.

In the evaluation phase the indicators used to some extent assume using solely one register. A more elaborated way of assessing the quality of modeling and linking in several steps would be useful. Also guidelines on how to measure the choice made and how to measure problems in the implementation phase could be useful. In the case of the sheep, the agricultural census can be used as a quality check in the future. However, it might be difficult to check the quality when the survey has been abandoned. Indicators for evaluating the quality in the long run as well as exit-plans might be needed.

The two frameworks used are similar. One advantage of the framework of Daas et al. is the focus on the procedures for accessing data and the legal prerequisites. Daas et al. also elaborates the output data quality to a larger extent than Laitila et al. The advantages of Laitila et al. are the focus on the production process indicators as a separate step and the discussion on how to evaluate the statistics by the quality criteria.

It becomes obvious that the quality frameworks can be used in the part of the process where a choice is made of whether or not to use the administrative register. The greatest usefulness of the systematic approach was in evaluating the administrative registers themselves.

**References**

[1] 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.

[2] Holmberg, A. (2012), Discussion on assessing quality of administrative data. *Statistica Neerlandica, 66: 34–40*. doi: 10.1111/j.1467-9574.2011.00507.x

[3] Zhang, L.-C. (2013), Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica.* doi: 10.1111/stan.508

[4] Mintzberg, H., Rainingham, D., Théorêt,A. 1976. The structure of unstructured decision processes. *Administrative* *Science Quarterly 21 246-275*.

[5] 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.Holmberg, A. (2012), Discussion on assessing quality of administrative data. Statistica Neerlandica, 66: 34–40.

[6] Laitila Thomas Wallgren Anders Wallgren Britt Quality Assessment of Administrative Data Researh and Development – *Methodology reports from Statistics Sweden 2011:2* Statistics Sweden.

[7] Wallgren, A., & B. Wallgren, 2007, *Register-based Statistics: Administrative Data for Statistical Purposes.* Wiley Series in Survey Methodology (New York: Wiley)

2.0 cm

1. (EC) no 1165/2008 and EC 1166/2008. Regulation (EC) No 1165/2008 [↑](#footnote-ref-1)