

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES
(EUROSTAT)**

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE/Eurostat Work Session on Registers and
Administrative Records for Social and Demographic Statistics
(Geneva, 1 - 3 March 1999)

Session 1, invited paper

Maintaining the quality of the registers used in the Danish census

Prepared by Marius Ejby Poulsen, Statistics Denmark

1. Introduction

1. Since 1981 the census in Denmark have been based solely on administrative registers. Information is gathered from several different registers and linked together by the use of a number of identifiers. This ensures a total count of every variable that is needed to establish the Danish census.
2. The experiences over the years in using administrative registers for statistical purposes have shown that the continuous and heavily administrative use of registers is the most important process in keeping the quality of the registers. In the administrative use of registers it is possible to detect and correct errors. The more intense the use of the register is the more errors can be found and corrected.
3. In addition to what the administrative users do for the register quality the statistical users have to act to keep track of the quality of the registers. In this connection exchange of information is a key topic.
4. In the paper some experiences of quality assessment and maintenance of registers, in connection with the Danish census will be described, on basis of a presentation of concrete examples.

2. Does the content of the register cover relevant subjects in a way that suits the statistics ?

5. The registers have administrative purposes and are designed to fulfil administrative needs. Changes in legislation and in administrative processes can give problems in relation to the statistical use of the registers. Therefore it is important to follow closely all changes that affect the administrative register and check the influence on the statistics. Some times it is possible by negotiations to have statistical views on the content of or changes in registers reflected in the decisions. Other times it is necessary to find a way to handle introduced changes. In this section an example is presented.

2.1 Integrated Data Collection¹

6. The situation where it is possible, through negotiations and co-operation, to have the statistical view reflected in the content of the administrative register, is included in the concept: Integrated Data Collection. The following example concerns the collection of data on work places, which is used to classify the sector in which the employed people are working.

7. The definition of integrated data collection is:

When an administrative authority obtains information on behalf of another authority when collecting information for its own use, it is called *Integrated data collection*.

8. This means that an administrative authority collects information from firms, institutions etc. which it will use itself to perform its own tasks, while at the same time collecting information which it does not itself need but is to be used by another authority, such as Statistics Denmark. The two data sets are collected in one and the same integrated operation.

9. In a number of cases Statistics Denmark needs both information collected for - and for the exclusive use of - Statistics Denmark, which must not therefore be used by the collecting authority, and information which the administrative authority collects for its own use. In these cases the combined set of information will make up the data which is supplied to Statistics Denmark and will be contained in the relevant statistical registers at Statistics Denmark. This applies in the Workplace Project:

10. The aim of the Workplace Project is to produce a list of all workplaces and to assign all jobs to the correct workplaces. Information for the Workplace Project is obtained by integrated data collection using the information sheets which employers have to submit each year to the Central Customs and Tax Administration in respect of each employee. The information sheets contain details of wages, pensions, fees, allocations from funds etc. for the individual employee or recipient of public benefits, who is identified by his or her person number. The employer is identified by an unambiguous identification number (the SE number), which is supplemented in the Workplace Project by a code for the place of employment, if the employer in question has several establishments. The information sheet also provides information on period of employment.

11. The Workplace Project is implemented once a year in co-operation between the Central Customs and Tax Administration and Statistics Denmark. The Customs and Tax Administration sends out the uncompleted information sheets in November/December each year, together with guidelines for firms on data to be reported for the year. These forms are accompanied by workplace lists for firms that have several establishments. The workplace lists are produced by Statistics Denmark. They contain a list of the workplaces which Statistics Denmark has recorded for the firm in question as they appear in Statistics Denmark's business register. The lists contain the name, address and sector (sector name and code) of each workplace. Finally the workplace lists contain the code for the workplace, which must be indicated on the information sheet for each individual employee. The employer must update the workplace list with appropriate information on the individual workplace, should there be errors or inconsistencies in the list. The employer must also complete the workplace list indicating new establishments and others which have closed so that, after any corrections needed, the list contains complete information on all establishments which have been active during the course of the year.

¹ The section is a rewritten version of parts of chapter 4.3 in Eurostat and Danmarks Statistik (1995).

12. The workplace codes are collected for use by Statistics Denmark. Information is also collected on the period of employment of the employee during the year in question for the exclusive use of Statistics Denmark.

13. Statistics Denmark receives a copy of all data shown on the information sheet, both the information collected by the Central Customs and Tax Administration for use by Statistics Denmark alone and the information collected for the administrative use of the Customs and Tax Administration itself.

14. Information from the Workplace Project is used by Statistics Denmark for statistical reports. In the first instance it provides the basis for individual data in the personal statistics system covering labour market indications for a whole year (data included in a register called AKM) and for statistics relating to a single point in time, end November (data included in a register called RAS). The data base-material is provided by the basic information from the Workplace Project together with extracts from a number of administrative registers and statistical registers, the information from which is processed in order to indicate labour market involvement in the last week of November of the year in question for all persons resident in Denmark on 1 January of the following year.

15. Concerning the yearly census, the relevant information is related to the labour market status, which includes data on sector and occupation. The former is based on the Danish version of ISIC and the latter based on the Danish version of ISCO.

16. The question: *Does the content of the register cover relevant subjects in a way that suits the statistics ?* must consequently be answered with a Yes, at least when we talk about the Census. The relevant subject *sector* is, as described, collected through integrated data collection, which ensures that data on the working place of persons employed is continuously updated.

17. The advantages of integrated data collection can be summed up as follows:

- Costs are saved in the process of data collection
- The consistency of the information collected is assured
- Firms are only inconvenienced once

18. The fact that statistical information can be collected in conjunction with administrative data means that the administrative data are statistically enriched and qualified. The information is consistent and is therefore better suited to covering statistical needs. Integrated data collection also ensures that some of the traditional weaknesses in the use of administrative data for statistical purposes are reduced. It is often claimed that administrative data are difficult to use as a basis for statistics because they are subject to changes in legislation and changes in administrative routines and procedures. This disadvantage can be reduced to some extent by the formulation of the statistical data which supplement the administrative data in integrated data collection.

19. Integrated data collection also involves certain disadvantages however:

- The information collected for use by Statistics Denmark may be alien to the administrative authority.
- Checking the correct reporting of information for Statistics Denmark may be low in the priorities of the administrative authority that collected the information.

12. Collecting information for another authority may give rise to problems since the administrative staff are not familiar with the information collected for that authority. Instructions and guidelines may ease the problem but interpretation in cases of doubt may be difficult for the collecting authority. In the case of the Workplace Project some of these problems are passed on to Statistics Denmark, since some of

the information on workplaces contained in the information sheet is supplied direct to Statistics Denmark.

13. It is natural for the collecting authority to assign top priority to its own information. That is what it knows and what is of central importance to the work of the whole institution, whereas it has no direct interest in the information for Statistics Denmark. If the data collected for its own use do not meet the formal and legal requirements, it must contact the firm and ask for new and correct information. Error detection routines are integrated into the process and are known, whereas this cannot be said of the other information collected. In the Workplace Project, for example, it was found that verification of the statistical data at the tax authorities was not sufficiently effective in all cases. It is therefore of crucial importance that as many process details as possible be agreed between the administrative authority and Statistics Denmark, in order to ensure that errors in reported data are discovered and corrected as quickly and effectively as possible.

3. Does the register give total coverage ?

14. One of the great strengths of register-based statistics, as compared with traditional questionnaire-based statistics, is that it is often possible to achieve total coverage, or something very close to it. One of the possibilities to check the coverage in the administrative registers is to compare the results with results from other sources.

15. This section contains a description of the process of linking and checking information from the Central Population Register (CPR) with information from birth- and death certificates.

16. Danish birth statistics relate to children born to women who are resident in Denmark. Statistics on births are based on two sources; information from the CPR, and information from the midwives' reports.

17. The CPR includes civil information on the child and the parents, all identified by their personal code number (person-number)

18. For each child born the midwives draws up a report with medical information. The child is identified by the person-number of the mother and by date of birth and sex (and number in birth in multiple deliveries). Almost all children in Denmark are born in hospitals and even when a child is born at home, there should be a midwife present. The midwives' reports goes to the National Board of Health and from where to Statistics Denmark.

19. Statistics Denmark links the information from CPR and the information from the midwives' report by using the person numbers and the information of birth-date and sex. This gives the opportunity to combine civil registration and medical registration and at the same time it is possible to get a check on the coverage of the register and of the system of midwives' reports.

21. The information from CPR and midwives reports are normally matched month by month. After a check of the validity of the person number the result of the matching process is:

- a) Births registered both in CPR and by midwives' reports.
- b) Births registered only in CPR.
- c) Births registered only in the midwives' reports.

22. Typically 99.2 - 99.4 percent of the births are correctly registered in both sources.

23. The non matching births are taken out for manual control. The most typical reasons for no match are:

- The midwives' report is missing (33 pct.)
- Registration in midwives' report of births where the mother is not resident in Denmark (5 pct.)
- The date of birth is not the same in the two sources (17 pct.)
- The registration is missing in CPR or delayed in order to be included (5 pct.)
- There are double registrations in CPR (10 pct.)
- There are double registration in midwives' reports (13 pct.)

24. The results from this process show that the quality of the information both in the population register and in the midwives' reports is very high. Under registration and double registration can be found more frequent in the midwives report than in the CPR, but still it concerns less than 1 pct. of the births.

25. The statistics on deaths comprise deaths occurring among residents in Denmark, dead in Denmark or abroad. Information on deaths are treated in very much the same way as the births. The two sources for death statistics are the CPR and the original death certificates. Like for the births the information from the two sources are matched by using the person number and the date of death.

26. Typically 99.3-99.6 percent of the deaths are correctly registered in the CPR and on the death certificates. Most of the non matching cases are due to the fact that around 0.4 percent of the deaths occur abroad among people resident in Denmark. In these cases there are no Danish death-certificates and the deaths are only found in the CPR. Besides people dying outside Denmark there are a small number of missing certificates and missing registration in the CPR. Like for the births there are a few cases of different date of death in the two sources.

27. As for the births we can conclude that the quality of the information in the CPR is very high.

4. How can the validity of data be checked ?

28. Linking registers together is one possible way to check the validity of data, for example, if different administrative registers contain redundant information (tax register and register of social benefit payments). Another possibility is to link register data and sample data dealing with the same subjects at micro level. In the following sections some Danish experiences on the last-mentioned are presented.

29. In Statistics Denmark the major use of administrative data is considered sufficient to cover most of the statistical areas, but of course not every corner of society can be described, which makes it necessary to carry out a few sample-based surveys.

30. Like the register-based data, the sample data are based on the person number. This enables the combining of data on micro basis (individual level).

31. The case where register data and sample data, within the same area are available, represents a straight forward opportunity of assessing both data quality and the methodology behind the production of statistics. In addition, when the data from the two sources are unambiguously identifiable, the assessment can be done on a micro level, evaluating individual data.

32. The study is about micro-matching of sample data and register data in the labour market statistics, with the aim of detecting discrepancies in similar² statistics from the two sources. The study was carried out in connection with a contribution to a report from the UN-work session on Statistical Data Editing, *Statistical Data Editing, Volume No. 2, Methods and Techniques*, UN, New York and Geneva, 1997.

33. It is obvious that the methods and principles used in the collection of data on sector, are essentially different in the two systems, mainly because of the fact that the sources of information are different. As a result, it would be surprising if the resulting statistics were consistent. What is meant by consistent is that every individual is classified equally in the two systems, e.g. according to sector. The results confirmed this assumption.

4.1 Comparing data on sector in register-based and sample-based statistics

34. In the study data on sector from the register-based labour force statistics (RAS, see section 2) and the Labour Force Survey (LFS) was compared. First of all the rate of inconsistency between the data in the two systems was estimated. The discrepancies were then analysed/distributed according to different characteristics, where the main target was to highlight areas where discrepancies were over the average.

35. One of the crucial problems in the comparison was the existence of a difference in the reference period. This was taken care of, by establishing a so-called stable population, as a sample of the total LFS-population, where the reference period could strongly be assumed the same.

36. Comparing industry codes at one-digit level (i.e. main industry groups) for the stable population gave the result that 90.3 pct. of the individuals had the same code in the two systems. Consequently 9.7 pct. had different industry codes.

37. One major problem when comparing the data in question, is that neither the LFS-data nor the RAS-data are without errors. Consequently, there is no checking-list. In the study it was assumed that the industry code was correct in RAS, in order to look more closely at the 9.7 pct. discrepancy to highlight on possible error-sources in the sample-based data.

38. The first error-source investigated was the *interviewers* carrying out the interviews. The result was that the more interviews the better result.

39. Another error-source was the *type of interview*, telephone-based or questionnaire-based. The result was, surprisingly, that the responses in the questionnaire-based interviews showed a higher rate of consistency.

40. A third factor which was analysed was the *type of respondent*, where the respondent himself, the respondent's spouse or finally other persons in the household could respond to the questions in the LFS. The result was clear: The information on which industry the respondent is working in is best given by the respondent himself.

41. The study presented above was carried out in order to document possible error-sources in the statistical data used in the labour market statistics. As mentioned in section 2, data from the register-based source (RAS) is used in the yearly census, concerning the classification of persons in employment. The results document the differences that occur when data are collected in two different ways.

² Similar means that both sources are used to measure the same statistic or variable.

42. In general the study represents a straight-forward way to check the validity of data. On the other hand, the interest might be claimed as more theoretical than practical applicable. It is obvious that if the staff resources available in the statistical bureaus were less limited, more evaluation studies would always be preferable. However, the marginal costs of using more resources should in this connection be taken into consideration.

5. Is the time reference in different registers the same?

43. In section 4 we saw an example of the problems encountering, when time references are inconsistent. When the Danish Census (and other statistics) are based on linking register information, the problem of time references in the registers is also present.

44. The registers used in the Danish Census are:

- the Central Population Register (CPR)
- the Building and Dwelling Register (BDR)
- the statistical Register of Employment Statistics (RES)
- the statistical Register of Education and Training Statistics (RETS)

45. In diagram 1 the time reference of each of the registers mentioned is illustrated:

Diagram 1: Time references of data in the Danish Census.

Register	Data	Time reference
CPR	Population, age, sex, marital status, etc.	1 January year x
BDR	Data on place of residence	1 January year x
RES	Employment status	The last week of November year x-1
RETS	Educational status	1 October year x-1

46. As it appears, demographic data and data on dwellings, refer to 1 January. Data on employment status (occupation, sector) refer to the last week of November and data on educational status to 1 October. Concerning the former the results would presumably be different compared to 1 January, dependent on the economic development. The latter however, the educational status of the population, must be assumed to be reasonably constant, due to the structure in the educational system.

47. All though there is these differences in time references it should be mentioned that in all cases the population is delimited as the population resident in Denmark 1 January. This ensures consistency in most of the information used in the Census, except from a few persons that change educational status or job in the period from 1 October and last week of November respectively to 1 January.

6. Concluding remarks

48. In this paper some examples have been presented, concerning assessment and maintenance of quality in the registers used in connection with the Danish Census. The first example underlined the importance and benefits of co-operation between the data suppliers - in most cases administrative authorities - and Statistics Denmark. The second example showed how alternative information about births and deaths can be used to update the population register. Finally, the third example described the possibility of linking sample based data and register-based data, on individual level, in order to check measures of the same topic.

49. The examples only represent some of the possibilities, when the statistical system is based on administrative registers like it is done in Statistics Denmark. In general it should be underlined, as mentioned in the introduction, that the continuous and heavily administrative use of the registers, is a significant factor, in respect of maintaining the quality of the information, which is later used for statistical purposes.

References

- Eurostat and Statistics Denmark (1995): *Statistics on persons in Denmark - a register based statistical system*, statistical document.
- Poulsen, M. E. (1995): *The LFS and the register based labour force statistics - a quality assessment*, contributed paper to the SMPQ-conference, Bristol 1-4 April 1995.
- Statistics Denmark (1992): *Statistiske efterretninger - Arbejdsmarked (1992:20), Arbejdsstyrkeundersøgelsen 1991*.
- Statistics Denmark (1993): *Statistiske efterretninger - Arbejdsmarked (1993:17), Registerbaseret arbejdsstyrkestatistik ultimo November 1991*.
- UN (1997): *Statistical Data Editing, Volume No. 2, Methods and Techniques*, UN, New York and Geneva, 1997.
