Distr.
GENERAL

CES/SEM.40/20
7 September 1998

Original: ENGLISH

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE-EUROSTAT work session on Population and Housing Censuses [1]
(Dublin, Ireland, 9-11 November 1998)

Study topic 1

THE POPULATION CENSUS 2000 BASED ON ADMINISTRATIVE SOURCES QUALITY ASPECTS

Supporting paper submitted by Statistics Norway [2]

1. Introduction

1. The *population* census in Norway for the year 2000 will be entirely based on registers, while the *housing* census will be a traditional survey. This paper deals with the population census. Information on the housing census is given in the paper "Establishing a dwelling register in Norway: The missing link."

2. The registers used in Statistics Norway (SN) are so-called statistical registers. The input data for all statistical registers in question are either administrative registers or other administrative sources.

3. Registers have been used by SN in collecting census data since 1970 [3]. In the 1990 census, information on the topics demography and education, as well as geographical characteristics, was collected from registers. Among the economic characteristics, data on income were register based. Labour market data was the only part of the population census where a survey was used. For

1 The papers which are prepared for this work session will be treated in the same manner , as papers that are prepared for seminars.

2 Prepared by Coen Hendriks and Paul Inge Severeide, Division for Population and Housing Census.

3 A brief history of the use of registers in Norwegian censuses is given in the paper "Establishing a dwelling register in Norway: The missing link."

persons living in small municipalities, the labour market data were collected by questionnaires only. For large municipalities a sample survey was used, and survey data were combined with register data.

4. For the 2000 population census, SN has decided to use administrative data for *all* variables, including labour market characteristics. This decision is based on the fact that the data situation has improved in several ways since the 1990 census. A brief description of the Norwegian system of register data is given in chap. 2.

5. The main theme of this paper, however, is quality aspects of register-based census data. Quality aspects are being described along four lines:
- Improving the quality
- Consistency between data from different sources
- Developing quality measures
- Quality control survey

2. A brief description of the Norwegian register data system

Population register

6. The existence of a Central Population Register (CPR) with a unique National Identification Number, has been very important in developing register-based statistics in Norway. This personal identification number is now used in all registers of any importance, and the ID-number is used as a linking key when matching registers. The CPR is the central register in population censuses.

Registers of activities and income

7. Several registers contain information on activities and income. The most relevant activity data in population censuses, are data on employment (job registers) and education. These registers are identifying the relations between the person and his/her place of work or school. They also contain additional information on activities, like ‘Occupation’ and ‘Hours worked’ for employed persons and ‘Field of study’ for students. Registers on unemployment and social security benefits give additional information on activity status. In addition, tax registers give information on different kinds of income.

Business registers

8. For employed persons, information on place of work, industry and type of sector is collected from business registers. A central register on legal units with unique ID-numbers was established in Norway in 1995, and most relevant registers are now using these ID-numbers.

The data situation has improved

9. Administrative records and registers have been used as sources for social statistics in Norway for many years. However, the amount of administrative

data available has increased over the last 10 to 15 years. The main reason for this is of course that administrative registers have been established in several new areas. Furthermore, an increasing part of the registers gives data on individuals (persons).

10. Until recently, most social statistics in Norway were based on single registers, or at least on combinations of data from registers belonging to the same administrative system (for instance tax registers). The increasing amount of administrative data available, and the fact that data from different systems can easily be linked, makes it possible to develop integrated statistics covering several areas.

11. In the 1990 census, SN found it necessary to use a survey to ensure the quality of labour market data. For the 2000 census, the decision is to rely on register data for all topics in the population census. The reasons for this decision are:

- The quality of the main registers has improved
- New administrative sources are available. (This includes establishing new registers and
- extending existing registers by new variables.)
- Data from several registers are being combined in an integrated data system to improve the
- quality of core variables
- SN has more information on data quality

12. This is not to say that the quality of all relevant register data is satisfactory today. Improving the register data is therefore an important part of the census project.

3. Improving quality of data

13. SN produces annual, register-based statistics for most of the topics covered by the population census. The quality aspects are handled by the ordinary production systems. In addition, as a part of the census project, extra resources are used to improve the quality of data from selected registers. In this way it is possible to include new variables based on registers in the statistics.

Improving the quality of each register

14. The quality of each of the administrative sources used, is of course crucial for the quality of the total population census. All registers contain some incorrect information. Coverage problems is one major type of error. This means that units are missing or wrongly included in the register. Furthermore, data for some register variables may be missing, inaccurate or misleading.

15. The main role for SN is to identify these errors and inform the administrative authorities responsible for the registers. This is of course an important activity in the ordinary production system in SN. As part of the census project, some extra resources are used to identify and correct errors,

thereby improving the quality. Errors may be more easily detected when we match data from different registers.

Improving quality by combining data from several registers

16. By combining data from several registers, we can handle quality problems in a better way. Combined data are especially useful in handling coverage problems, but may also be used in correcting missing or inaccurate data. An example of this is matching labour market registers and registers of different types of income. By using the information from all these registers simultaneously, the quality of variables describing the labour market situation, as well as income variables, may be improved.

New variables by matching registers

17. Some of the core variables in the population census are collected directly from administrative sources (e.g. sex, age). For derived topics like 'Socio-economic groups', many variables from different registers have to be used. But even for other variables, like 'Current activity status' (employed, unemployed, not in the labour force), data from several registers have to be combined.

18. We will use the variable 'Current activity status' as an example. The main problem in this case is to identify employed persons. There is no single register covering all groups of employed persons. The main registers used are:

1. Register of employees
2. Register of unemployed persons
3. Tax register of wages per year
4. Tax register of income (self employed persons)

19. According to the ILO definition, persons should be classified as employed if they work for at least one hour during the reference week. This is the definition used in the Labour Force Survey (LFS). In the population census, as well as in the annual, register-based, labour market statistics, the goal is a similar classification based on register data. This system of classification of employment, will use information from all the registers listed above. The classification rules will be rather complex, and only the main principles are described in this paper.

(1) The only register that has reliable information on duration of jobs, is the *register of employees*. A person having a job in the reference week according to this register, will normally be classified as employed. The exception will be when no wages are reported in the tax register.

(2) The information on duration of jobs in the tax register of wages, is less reliable. In most cases the duration is reported to be the whole year. Persons with jobs only according to this register (and not in the register of employees), in most cases have to be classified either as employed or not employed for the whole year. The most reliable information is the total wages for the year. Persons with wages exceeding a certain limit, will generally be

classified as employed. The wage limits will be computed by using information from the LFS.

(3) For self employed persons, we have no information on job duration at all (only income for the whole year). The classification will be carried out as in rule (2), but we use ‘Income as self employed’ instead of wages.

(4) Persons who are registered as unemployed in the reference week, will normally be classified as not employed (more details in chap. 4).

20. The quality of the resulting classification has been analysed by using data from the LFS at a micro level, and the classification routines have been adjusted. The final result is that 85 % of all employed persons are classified by routine (1), 7 % by routine (2) and 8 % by routine (3).

Using surveys to complete register data

21. In some cases, it is necessary to complete register data by using surveys. The register of education is not updated with education taken abroad. As a part of the 2000 census, a survey will be carried out to collect information from persons educated abroad. A similar survey was carried out as a part of the 1990 census.

22. Another example is the use of information from the LFS in constructing the labour market variables. As mentioned above, these data play an important role in the classification of current activity status. LFS data will also be used to improve the quality of other labour market variables, like ‘Time usually worked’ and ‘Occupation’.

4. Consistency

23. Obtaining consistency between variables is an important activity in any population census. This is indeed the situation when the census data is collected from several administrative sources and not from a questionnaire. By consistency between register data, we mean that linking information from different registers should give consistent classifications of activities for all persons.

24. The implementation of this general definition is however far from simple. Some combinations of activities are obviously not logical and can not be accepted. A person can not be classified as employed and unemployed at the same time. Other combination may be “legal”, but are not very likely. Some combinations are, on the other hand, very common (many students have a job). The main point is to determine what are real combinations of activities and what are inconsistencies between register data.

Types of inconsistencies between register data

25. It is useful to examine different types of inconsistencies more closely, since they may have different causes.

Different definitions of variables

26. What is in fact inconsistency between register data, depends on the definitions of the variables compared. The formal (or legal) definitions of the different variables must be compared. When a person receives wages according to the tax register for a period, this indicates that the person should be registered as an employee. But according to the criteria for registrations in these two administrative systems, there are several exceptions. In other words, consistency can not be expected for all persons involved.

Different populations

27. Use of different populations may cause inconsistency. This is not a very large problem in population censuses. The problems are much greater for business registers where the units are establishments or enterprises.

Differences in reporting routines

28. The formal definitions decide what *should be* the contents of a register. The real contents will also depend on the implementations of these definitions in the administrative system. This underlines the importance of studying the reporting routines as a part of the quality control.

Different reference periods

29. One important reason for inconsistency is the use of different reference periods in the registers involved. The information collected may refer to a specific day in some registers (but not necessarily the *same* day in all registers), or a longer period like one year in other registers.

Errors

30. In addition to the types of inconsistencies mentioned above, all registers may contain incorrect information.

An example: Inconsistency between the register of employees and the register of unemployed

31. The potential of handling inconsistency between register data, will be presented by an example based on the linking of the register of employees and the register of unemployed persons.

32. Since both registers use the same personal ID-number, the linking creates no problems. The number of persons with a non-valid ID-number is very small. The real problem is to produce consistent data for all persons involved. This is done by correcting the duration of either the job or the unemployment period when these periods overlap.

33. According to ILO definition, persons can not be classified as unemployed and employees at the same time. However, we found that for the year 1992, 33 % of all unemployment periods overlapped with one or more jobs.

34. Even though it is against administrative rules, a few persons may be registered as unemployed at the Employment Office even if they have a job at

the same time. In most cases, however, inconsistencies are caused by errors in one or both of the registers. There are two main explanations:

1. Job termination may not have been reported to the register of employees, or the date of termination or start reported may be incorrect

2. The date of leaving unemployment may be inaccurate in the register of unemployed persons

35. Based on studies of the reporting routines of the two registers, we have laid down some rules for correcting the duration of the periods. The main criteria for correcting the periods are

- What period started first (job or unemployment period)
- The duration of the overlapping period
- Whether the job is registered as terminated or not

36. Generally we correct the duration of the job when the overlapping period is long and the unemployment period when the overlapping period is short. When the job has not been reported as terminated, the duration of the job is always corrected (duration of overlapping period cannot be computed in this situation).

37. The effect of using this correcting routine would be to reduce the number of employees for a reference day (the date chosen was December 1, 1992) by 3,4 %, and the number of unemployed persons by 2,0 %. The routine is yet to be evaluated, and has so far not been used in producing labour market statistics.

5. Measuring the quality

38. As stated above, improving the quality of the register based data, is one main issue in the census project. Equally important is to produce and present information on the quality of the data involved.

Measuring the quality is important in two ways:

- Monitoring the quality as a part of the production process
- Describing the quality of the statistics published

39. Each of the register-based statistics that are currently being published by SN, are followed by some kind of quality description. The aim of the census project is to develop standardised ways of describing the quality of all variables contained in the census.

40. Most statistical quality measures are designed for surveys. An important part of the census project will be to develop quality measures especially suited for register-based statistics. The intention is to use these methods for other register-based statistics as well, not only for census statistics.

6. Quality control survey

41. In order to measure the quality of the census , additional data sources are required. Population censuses are traditionally followed up by a quality control survey. The purpose is to collect data on selected variables in order to measure the quality of the corresponding census data. Data may be collected in different ways:

1. Independent sample survey
2. Additional questions to existing surveys
3. Data from existing surveys

42. Traditionally, quality control surveys have been carried out as independent sample surveys. In the 2000 Census the plan is to use alternatives 2 and 3. The fact that the census is based on register data rather than a questionnaire, makes it possible to use data from existing surveys more efficiently. We can use the Labour Force Survey (LFS) as an example.

43. The core variables in the field of the labour market, are related to the current activity status. This means that we collect data on the labour market situation for each person for one reference week. The same kind of data is collected in the LFS. Hence the LFS may be used as a source for quality control for the labour market variables of the census. The number of persons interviewed in the LFS in one week is not more than 2000. However, using registers, “census data” may be established not only for the census week (first week of November 2000), but for all weeks in a longer period, for instance the last quarter of 2000. This gives us the possibility of using data from 24 000 interviews in the LFS.

44. For variables describing ‘Usual activity’ (referring to the situation in the year 2000), the LFS cannot be used directly as a data source. The alternative in this case is to ask some additional questions in the LFS in the first quarter of 2001.

45. We estimate that these methods of data collection will reduce the cost of the quality control survey by approx. 30 % when compared to an independent survey.

7. Conclusions

46. The decision to have a totally register-based population census 2000, is grounded on the fact that the quality of the register data has improved over the last year. However the quality is not yet satisfactory, and it is considered as an important part of the census project to focus on quality aspects. This involves improving the quality of selected variables, combining data from several registers, obtaining consistency between data from different sources and standardising quality measurement and description. If these quality projects turn out to be successful, SN will be able to publish a “population census” annually in the next century.