LINKAGE OF RECORDS ON OBJECTS OF DIFFERENT KINDS. METHODOLOGICAL PROBLEMS AND PRACTICAL EXPERIENCE.

Poul Jensen
and
Lars Thygesen

Danmarks Statistik
Copenhagen, Denmark

1. INTRODUCTION

Since the late 1960s, much effort has been devoted to improving and extending the use of administrative records for statistical purposes. It has been generally agreed that the key to progress in this field is record linkage. This is because the information on citizens which is gathered for the purpose of running one branch of public administration, e.g. social security, is not sufficient for compiling socio-demographic statistics, so the data must be supplemented from elsewhere.

The general approach of statisticians has been to regard the administrative records as a supplement to survey or census data, and the purpose of linkage procedures has mostly been to put together data concerning one person from a questionnaire and from some administrative file, thus enriching the survey information. As a consequence, the problems of linkage of records on the same persons have been much discussed. In countries where a unique and permanent identification number for persons, a Person Number, has been introduced throughout the public administration the technical problems of such record linkage tend to be small. Other problems related to the interpretation of the combined information may be serious enough.

This technique of linking records from different sources relating to the same persons is widely employed in the Danish system of socio-demographic statistics, which rely almost exclusively on administrative records, the linkage being facilitated by the existence of a Person Number.

However, linkage of data on different objects (persons, dwellings, enterprises, etc.) is sometimes required to make full use of administrative records. For instance, it is desirable to be able to link individual records of family members to create new information on the family; and information on persons from a population register should be combined with information on dwellings to create housing statistics. Linkage of this kind tends to create greater methodological problems, and these problems are the issue of the present paper. The problems will be discussed on the background of two cases of Danish experience, which are presented in sections 3 and 4. Before that a very brief account of the Danish statistical system is given in section 2. The issues of privacy and data protection in relation to record linkage will not be discussed here.

2. THE COMPILATION OF STATISTICS IN DENMARK

The compilation of statistics in Denmark is more centralized than in many other countries. The central statistical office (Danmarks Statistik), which is responsible for most social statistics of a general nature, is an independent public body which was set up under a special law. Danmarks Statistik is entitled to collect administrative data from public authorities for statistical purposes.

Since the beginning of the 1970s, one of the aims of the strategic planning undertaken by Danmarks Statistik in the field of statistics relating to persons, has been to set up a coherent system of statistics based on information from administrative registers, seeing that public administration in Denmark relies rather heavily on recorded information relating to the "objects" of administrative action, e.g. citizens, business enterprises and buildings.

2.1 The administrative registers

Since 1924 every municipality in Denmark has kept a local population register, i.e. a file containing information about all persons living in the municipality. These registers contain identifying information such as occupation, name, birth date and place of birth. Apart from these, the most important items of information are the address, family circumstances and nationality.

The municipalities continuously update the files, using information on births, marriages, deaths, etc., obtained from various public authorities. The individual citizens are obliged to report any changes of address directly to the registration office.

A major reform of the population register system took place in 1968. The municipal registers continued to exist, but in addition a central population register, CPR, was created. This is a computerized register covering the whole of the Danish population. The central register and the local registers are updated as part of one coordinated administrative process.

An important part of the reform was the introduction of a permanent and unique identification number for every citizen: The Person Number. This number was regarded as a practical necessity for the operation of the central population register. In addition, it was to be introduced in every area of public administration, thus replacing the different number systems which had hitherto been used by the various administrative departments.

Information from the CPR is used by the public administrative bodies in almost all areas relating to the individual citizen. This means that there are many opportunities to identify and correct or remedy errors and defects in the information contained in the register.

In the years following the creation of the CPR, the use of computers by Danish authorities greatly increased, and large registers of persons were created to administer the collection of taxes, the payment of pensions, etc. All these registers, which are valuable sources for statistics, use the Person Number as an identifier.

In 1977 a law was introduced which set up a nationwide buildings and dwellings register to be used by the municipal authorities. As explained in section 3, the prospects of using the records for producing census statistics contributed to the decision of establishing the register.

Finally, mention should be made of the Central Register of Enterprises and Establishments, which contains basic data on both enterprises (legal units) and establishments (local units). The register, set up under an Act of Parliament of 1975, is kept by Danmarks Statistik.

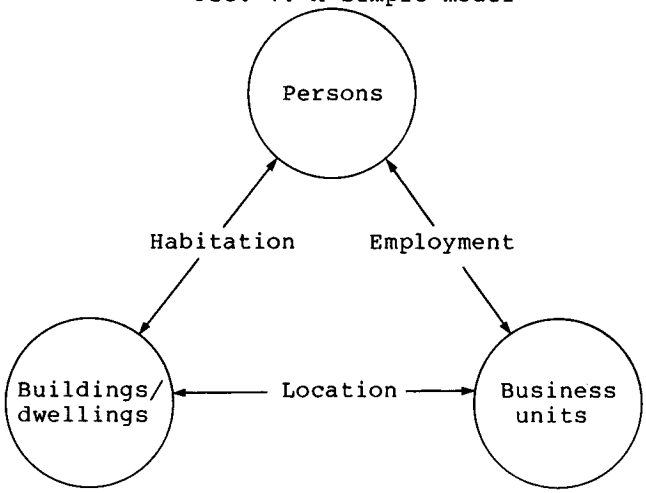
2.2 The principles of the system of socio-demographic statistics

The Danish system of socio-demographic statistics has developed progressively since 1970 in parallel with the creation of the administrative registers on which it is based. The first step was to reorganize the annual vital statistics, based on the CPR, comprising tabulations of the population and its movements, with information broken down by sex, age, place of residence (municipality), etc. Statistics on income, employment, etc., then followed. The basic development work was not completed until 1981.

The system's effectiveness was demonstrated by the 1981 Population and Housing Census which was carried out without sending questionnaires to the public, but solely by collecting information already available in the system. This kind of census could, in principle, be carried out every year.

The statistical system is concerned mainly with persons. It also contains information on the dwellings where these persons live and on the business units where they work. In the system, these three types of objects are linked by means of the unique identifiers for each subject: person number, the exact address of a dwelling and a code number for the workplace, as shown in fig. 1 below.

FIG. 1: A simple model



Each object has a number of characteristics, such as age, occupation, and number of rooms in the dwelling, which constitute useful statistical information regarding the social situation of the persons.

The aim of the system is to create a statistical description of the persons' social and demographic situation and of changes in this situation. The description is intended to enable many types of statistical analysis to be made.

The basic information on the three objects mentioned (persons, dwellings and business units) and on the relationship between them is contained in a number of independent statistical registers. Each of these registers is designed to be used for a single statistical area (e.g. education statistics) and contains the data required for this purpose. At present there are 37 statistical registers. Two or more registers are combined to extract anonymous statistical data only if this is required for an actual study.

Most of the registers are updated once a year although some are updated more frequently. The updating work normally consists of a statistical revision of extracts from one or more administrative registers. The basic information is compared, checked and amended so as to obtain the best possible overall estimates of the relevant characteristics. The sources for a statistical register often consist of more than one administrative register because the combinations of data which occur in the administrative registers are not relevant to those who use the statistics.

Since the linkage of data using person numbers is an essential part of the updating procedure, information extracted from the source registers must contain personal identifiers.

Part of the updating procedure is to establish and validate the links between persons, dwellings and business units.

These links cannot be established through the use of common identifiers. A key containing the identifiers of both types of object must be etablished. This is the topic of the following sections.

3. LINKING OF DWELLINGS AND INHABITANTS IN REGISTER-BASED HOUSING STATISTICS.

3.1. The background

Acting on the advice of Danmarks Statistik, the Danish government decided in 1974 that a mid-decade traditional census in the term 1975/76 should not be taken. The decisive argument in the recommendation was the long-term advantages that were supposed to be gained by further efforts as regards the development of register-based census-type statistics. An explicit precondition for the census decision was that a Central Register of Buildings and Dwellings ("Bygnings- og boligregistret", abbreviated BBR) had to be established.

Such a register had been under consideration for some time, mostly for administrative reasons. However the issue was brought strongly forward in advance of the census decision, notably from the local governments. One of their main arguments was the wish to relieve the public and themselves of the burdens of a traditional census.

The responsibility for the organisation of the BBR was placed with the Ministry of Housing, whereas the local goverments were responsible for the setting up and updating of the registers. According to legal provisions Danmarks Statistik is under an obligation to contribute to the planning and utilization of administrative edp registers which can be used for statistical purposes. Since this undeniably was the case as regards the BBR, Danmarks Statistik had to accept the challenge: The question was not if it was possible to compile register-based housing statistics, the question was alone how this should be done.

3.2. The basic problem

In most traditional population and housing censuses the information on persons and dwellings is placed on the same questionnaire, and the two types of information are linked to one another through the compilation processes as long as this is desired. The object type "person" and the object type "dwelling" are connected by an "object relation", which in this case may be symbolized by a questionnaire number. This automatically existing object relation implies that housing statistics, compiled on this basis, not only may show the characteristics of dwellings as physical objects or units but also the interrelation of the characteristics of the dwellings and the number and properties of the occupants. This feature of course is essential for the housing statistics.

Given the existence of one register (CPR) with information on the persons and another (BBR) with information on the dwellings, it is of course technically possible without major methodological problems to compile statistics referring to each of these objects separately. On the other hand it is by no means a matter of course to establish the basis for combined housing statistics. The precondition is that the object relation - since it does not exist at the data birth as in the censuses - may be subsequently established. In other words: the records in the BBR concerning dwellings must be linked to the records in the CPR concerning persons.

This of course depends on the existence of some sort of common denominator. One way to construct such a facility could be the establishment of a third register containing records which combine the identifiers of the two registers. A solution along these lines - but in another field - is described below in section 4. Another possible solution would be to use a common (secondary) and parallel updated identifier.

It was this last method, which was used in the field of housing statistics. The point of departure in the search for an effective common identifier was something rather earthbound: the postal address.

3.3. The original idea

In Denmark the general, traditional postal address has the following composition:

(Name of recipient).
Street name.
Street number + letter where needed (identifying houses/entrances).
Floor designation where needed (a combination of special indications and numbers).
Apartment indications where needed, typically indications as "to the right", "to the left", "middle" etc.
Postal district.

When the BBR was planned, street names and numbers were not in use in rural areas where the postal districts then were very small, but they were being planned in consequence of the large-scale amalgamation of municipalities (local government reform 1970) and of postal districts.

The CPR contained the postal address of the domicile of the registered persons in a systematic, but originally somewhat incomplete, way. However, it was now suggested, particularly by the large urban municipalities who have played a leading role in the endeavours to make the BBR effective, that the address should be developed into a common identifier for BBR and CPR, by means of general systematizing and formalisation of the system and especially by substituting the most inaccurate and ambiguous parts of the address designation - floor number and apartment orientation - with specific numbers.

The idea was that a sign or number plate with these numbers should be placed in front of each apartment in houses or staircases with more than one apartment, in order to bring about unambiguity in the address designations on the spot - and presumably in the registers.

However, this last idea was not carried through, since the Danish parliament decided against it, so the foundation of the record linking between the BBR and the CPR had to be secured in another way.

3.4. The solution

The final solution was based on two measures. The first was that the above-mentioned, already existing address-designation system was systematized, so that in every practical sense it had the properties of a formal coding system, even though to outsiders it did not appear so. To this end, rules were adopted for the sequence and exact designation of the various address elements. As to the distinction between apartments placed to the "left" or "right" - an issue which had been a matter of discussion and even ridicule in connection with the debate on the number plates - it was decided that the proper angle was the one of a person standing in the staircase and facing the apartment doors. This was a victory for the postman´s angle over the angle of the building authorities, whose tradition was to decide on the question of "right" or "left" from a place out in front of the building. It may be added that in many Danish buildings the two different angles will lead to the opposite results. Thus the question as regards general-purpose register-systems is of more strategic importance than it may seem.

The second measure was the establishment of a system constantly controlling the consistency between the address designations actually used in the CPR and the BBR respectively. This system was created along the following lines:

1) at a given stage in the developing process, the BBR and the CPR were matched on the address designations, and discrepancies were corrected by the local governments´ BBR units.

2) a file ("the address index") containing all existing correct addresses was hence constructed and integrated in the BBR. The updating of the address designations themselves was based on information from the building authorities of the local governments.

3) the updating of the actual addresses of the individual persons was to be based on the fact that persons in all instances where they take up new residence have to notify the local population register. From now on they were asked to give the complete and correct adress, which subsequently was checked against the address-index file. The local population register authorities were responsible for correcting the address if it was incomplete or invalid.

The aim of these measures was to prevent that errors were accumulated over time, making effective linkage impossible.

3.5. The results of the linking

The success of the address-linking system was considered a key problem in the construction of the BBR and generally in the Danish developing of register-based statistics.

After the careful planning the rather complicated address-designation system by and large succeeded, whereas other BBR problems - notably the updating of new buildings and dwellings - subsequently appear to have been more difficult to solve.

The decisive test of course was the compilation of the housing statistics, based on exact matching by means of the address designations from the BBR and CPR respectively.

The rate of non-matching records is shown in the following table:

Register housing census date	"CPR persons" without corresponding address designation in BBR (% of total population)
1/7-1979	4.0 per cent
1/1-1980	2.1 " "
1/1-1982	1.2 " "
1/1-1984	1.0 " "

Some of the differences are due to a timelag between the updating of BBR and CPR. Besides the figures vary according to the age and structure of the buildings and to the size of the municipality, but must by and large be considered satisfactory. Of course an unknown number of persons are wrongly "placed" in dwellings. In general, this type of error does not seem to be of substantial importance.

The statistics of occupied dwellings and households are based on those CPR and BBR records for which the matching is positive.

In principle it should also possible to estimate the number of vacant dwellings by counting the dwellings records which do not match with CPR persons. This is in fact not feasible, notably because the numerical difference between the total stock of dwellings and the occupied dwellings may be influenced by a number of factors, for instance the following:

(i) dwellings occupied by non-residents (for instance diplomats)

(ii) dwellings which (without being constructed as weekend-cottages etc.) are "secondary" dwellings

(iii) dwellings occupied by residents who in the CPR have erroneous but valid address designations

(iv) dwellings used exclusively as apartments for visitors, for instance of business companies

(v) dwellings which already have been let or sold, but so far not taken into use by the new tenant or owner

(vi) dwellings that have been demolished (or converted to other use) without the BBR so far having been notified (which is one of the weak points in the BBR updating).

Each of these categories may be small, but together they will draw a veil over the number of vacant dwellings ready to be let or sold.

This observation concerning the possible use of non-matching records as a basis for estimation of marginal quantities may have some general application. Such use, which is especially relevant in longitudinal statistics, seems to require very accurate information.

At any rate in the CPR/BBR context it has not yet been possible to realize some hopeful early ideas of gaining an important side benefit in the form of new general statistics of vacant dwellings.

3.6. The experience

The exposition above presupposes implicitly that the edp-technical problems of record-linking are small once the material foundation for the linking is established in the form of an effective identification key. Danish experience of linking BBR and CPR confirms this suggestion, even if the linking (which for statistical purposes is carried out by Danmarks Statistik) may be somewhat affected by the fact that the administration of the CPR and BBR registers is performed in two different public edp bureaus.

The experience drawn may thus be that the linking of records of dwellings and persons respectively is possible on certain specific conditions, cfr. above, and that this method is a most valuable new statistical tool. It may be noted that an identification system does not necessarily demand large administrative administrative register systems of the type used in Denmark.

Of course the final quality of the housing statistics depends not only on the effectiveness of the record linking, but also on the concepts, and on the validity and statistical relevance of the contents of the registers in question. However, this is not the proper place to pursue this far-reaching issue. It may be enough to note that the Danish CPR/BBR housing statistics compare favourably with the former traditional Danish housing statistics since they

- are more timely and frequent (annual or biennial)

- contain a larger and increasable amount of data (in fact nearly all other types of register data may be drawn upon, if necessary).

The statistical validity of the data is generally spoken acceptable or better. In the authors´ views it varies according to item from not quite satisfactory (some details in the BBR) to excellent (the major part of the CPR).

4. LINKING OF BUSINESS UNITS AND EMPLOYEES.

4.1. The problem

The Danish decision to base future population and housing censuses on administrative registers instead of questionnaires presupposed that it would be possible to fill a number of information gaps. Thus it was necessary to establish a link between data on employed persons and data on the business units where they work.

This link, for one thing, was required for the compilation of statistics on occupation by kind-of-activity. These statistics are traditionally produced as part of a population census and they are indispensable for planning purposes. It was also found necessary to establish statistics on commuting, based on a combination of information on the persons´ home address and work address. In addition, the link between persons and business units would be extremely useful in business statistics.

On the one hand we have data from the CPR and other person register containing information on the persons´ home address, occupation, etc. Each person is identified by a Person Number. On the other hand, as mentioned in section 2, information on local business units (establishments) is kept in the Central Register of Enterprises and Establishments, and an identification number is assigned to each of these units. But whereas the updating of records on legal business units (enterprises) is fairly reliable because these records are used in public administration, there is no administrative routine for updating records on local units.

4.2. The solution

First of all, the registration of existing local units had to be rendered complete. This was accomplished by an annual enquiry among employees.

Secondly, how could a reliable link between employees and work places be brought about? Part of the solution to this problem was found in the tax administration, because employers were liable to report wages and salaries etc. to the tax authorities for control purposes. These reports were gathered in an edp-register called the Salary Information Register, containg (in addition to the amounts earned) information on the identity of the employer and the Person Number of the employee. If the employer has no more than one workplace, the link between employer and employee is sufficient. Otherwise the records need to be augmented by a "workplace code number" if they are to fulfill statistical needs.

The tax authorities did not need this augmentation for their own purposes. But they accepted to extend their Salary Information forms by a box space to be filled in with a four-digit workplace code number (supplied by Danmarks Statistik) by the rather few employers with more than one establishment.

Though this extra information is collected and registered by the tax authorities, it may be used for statistical purposes only.

The described method of data collection was only applied to private employers, as a more direct reporting procedure could be created from public wage administration systems. The orders of magnitude in the project can be seen in below.

	Number of employers	Number of workplaces	Number of registered employments in Salary Information Register
Total private sector	197,000	220,000	3,700,000
Employers with one workplace	190,000	190,000	2,800,000
Employers with more than one workplace	7,000	30,000	900,000

4.3. Securing data quality

A weakness of the solution is that part of the data is not used in the administrative process. This leaves Danmarks Statistik with the heavy responsability for securing the quality of data.

The first problem is to persuade the 7000 employers to give the extra information for statistical purposes. Even though Danmarks Statistik according to law has the power to demand the information on workplace from employers, statistics tend to be taken less seriously than administrative control purposes. It is therefore essential that it should take as little effort as possible to supply the data. To accommodate them, employers are not forced to use the code numbers assigned by Danmarks Statistik on the individual reports; they are free to give internal "department numbers", used in their own information systems, as long as they explain the meaning of the numbers. Nevertheless there is still after five years of data collection a fairly large proportion of employers - around one third of the employers, most of them small - who fail to report the workplace code number in the original report to the tax authorities.

It is of course essential that such deficiencies, as well as other errors, are detected and returned to the employers immediately. An initial computer procedure in Danmarks Statistik finds different kinds of formal inconsistencies in the Salary Information Register, compared with the Central Register of Enterprises and Establishments. In case of errors, a report on the employer in question is produced and used as a basis for a telephone call to

the employers' wage administration. In most cases the errors can be identified and corrected during this telephone call, but sometimes it is necessary to produce a special form to be filled in by the employer.

The formal errors include the following kinds:

(i) An employer who is not known to have more than one establishment fills in the code number; often this indicates that there are in fact more than one establishment.

(ii) Missing workplace code numbers.

(iii) Use of code numbers other than those assigned by Danmarks Statistik or explained by employer; this may also indicate workplaces that are unknown to Danmarks Statistik.

Even if the reports of one employer do not offend against these rules, this does not mean that they reflect reality and that the linkage result will be correct. Therefore, a number of checks are made in order to pick out employers with a great probability of errors. These checks include comparison of employment figures from one year to another, and analyses of the variation of distances between the workplace and the homes of employees; if many of the employees seem to have to travel a long way to work, something is probably wrong. The following types of errors can be detected by these procedures:

(iv) The address of the recorded workplace may not show the actual position but for instance the postal address of an administrator.

(v) Two or more workplace code numbers may have been confused by the employer.

(vi) The same code number has been used for two workplaces, leaving one of them to appear without employees.

4.4. The result

The error checking and verification task is a "manual" process involving a considerable amount af work (five qualified man-years per year). Contrary to early expectations, the amount of work has not yet decreased. As a consequence it has so far only been possible to produce the results with a considerable delay, almost one and a half years after time reference of the statistics.

It must also be recognized that all errors are not found in the checking phase. Some of them are insignificant in relation to the use to which resulting statistics are put. But a number of more important errors are probably never found, and a few have also been detected after the release of the result, which is extremely undesirable. In such cases the checking procedure is reviewed to prevent similar situations in the future.

As a consequence of this and of a slowly growing data discipline among part of the employers, the reliability of the occupa-

tional and commuting statistics has been ever increasing, a fact which is recognized among local govenment users of the statistics. At the same time the project has improved the data quality of the Central Register of Enterprises and Establishments.

5. CONCLUDING REMARKS

If statistics are to make full use of administrative records, methods of linking data relating to different objects must be explored. This is always a difficult task because no common identifier exists. It is a great help if the administrative bodies creating the records need the linkage procedure themselves. Otherwise the statisticians are left with the burden of securing the quality and persuading respondents to clear up errors.

The idea of the Danish approach to this kind of linkage has been to use, as far as possible, existing elements of the administrative procedures. The adjustments that have been made are rather slight and have not been felt very much by the majority of "respondents".

In the authors´ opinion this has contributed to fairly successful solutions.

SUMMARY

Producing statistics by automatic linking of records of information is by now a well-established practice in several countries. The literature on this subject has so far been dealing largely with attempts to link information on one and the same object from different sources. The focus of the present paper is on describing situations where the aim is to combine information on different subjects. One example: Linking of information on a person (age, residential address, etc.) with information on his workplace, its address and kind-of-activity. Records linkage of this kind for statistical purposes can be extremely fruitful, and it is employed in the Danish statistical system, based on administrative records. But it involves a number of methodological and practical problems that have to be overcome. One group of problems concerns the creation of the key between the objects to be linked. Having established that link, methods must be designed to ensure the validity of the linkage; experience shows that mis-linking is likely to occur and rather difficult to detect.

The paper considers the outcome of two cases where linkage of records on different objects has been done in Denmark.