WP. 29
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE	**EUROPEAN COMMISSION**
CONFERENCE OF EUROPEAN STATISTICIANS	**STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)**

Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (v): Privacy for new types of microdata: sequence data and mobility data

ANONYMISED INTEGRATED EVENT HISTORY DATASETS FOR RESEARCHERS

Prepared by Johan Heldal, Statistics Norway

Abstract

Statistics Norway (SN) maintains a social security event history database called FD-trygd which is constructed by integrating a large number of registers maintained in various public institutions through exact matching. FD-trygd describes movements between states for a large number of event variables and economic variables associated with demography, social benefits, labour marked, pensions, education etc. for every person with residence in Norway since 1. January 1992. The data are of high quality and of large interest for research purposes.

In 2009 the Norwegian Social Science Data services (NSD) and SN agreed that NSD should be entitled to keep a 20 percent sample of all the event histories at its premises for distribution of anonymised data to researchers. The data should be made anonymous at NSD premises according to rules set by Statistics Norway.

This paper will describe the solutions SN has come up with so far and some challenges we face in risk assessment and anonymisation of such longitudinal data. A preliminary solution is now being tested among researchers. The outcomes of the tests will influence the further development of the anonymisation procedure.

1. INTRODUCTION

Dissemination of microdata to researchers has a tradition in Norway since the 1970s. Trusted researchers at universities or educational institutions established by national law and other approved research institutions and researchers working on projects funded by the Norwegian Research Council can apply for confidential[1] data from Statistics Norway. The datasets to which access is granted can originate from sample surveys carried out by Statistics Norway, from registers that SN commands for statistical purposes or both. The register owner and the Norwegian Data Inspectorate must approve the access.

Norwegian Social Science Data service (http://www.nsd.uib.no/nsd/english/index.html) was founded by law in 1971. One of its main responsibilities was to simplify access to data for researchers and students, included anonymised micro data from Statistics Norway. SN has the responsibility for the anonymisation of these data.

Researchers granted access to confidential data from SN as well as those receiving anonymous data from NSD have to sign a contract which involves professional secrecy. Through all the years that micro data has been disseminated this way breach of confidentiality has never been observed and the arrangements have provided rich research opportunities and a substantial added value of the micro data for the Norwegian society.

Section 2 gives an introduction to the social security database called FD-trygd from which NSD has requested and been granted a 20 percent sample for anonymisation to researchers at their premises. Section 3 describes the principles laid down by Statistics Norway as a basis for establishing an anonymisation procedure. Section 4 describes how SN has tried to implement these principles into a set of rules that NSD can use for test deliveries.

2. WHAT IS FD-TRYGD?

FD-TRYGD (ForløpsDatabase-trygd) is an event history database based on administrative data and contains all events that have taken place in relation to the Norwegian Social Security System for every individual who has lived in Norway since 1. January 1992. The event variables are categorical variables that may change value (state) at points in time such as marital status, employment registration status, sick leave, disability status, pension status etc. For every such variable the database has records showing all dates that these variables have changed and the new value valid from that date. Economic variables associated with states of the event history variables, such as unemployment benefit, social security benefits etc. are included. Data are kept in separate Oracle tables. The tables can be exactly merged using personal identification code. When this is done for statistical purposes the personal identification code is replaced by a serial number only to identify which records belong to the same individual. By merging tables we can construct a history file which for an imaginary female immigrating to Norway 22. November 1999 may look like table 1.

The variables “Sick leave” (yes, no) and “Maternity leave” (yes, no) are basically separate variables from employment status originally and occurring in different tables, but have here been combined into one employment variable with these categories. Sickness and maternity benefits to replace income from work have here been included in a separate column.

Considerable more complex histories occur or can be constructed. There is a large number of tables and variables in the system. Education and yearly income based on the tax assessment can be added although income is not an event history variable but relates to a year at the time. Essentially, the only constant variables in the system are date of birth, country of birth, immigration date and country of immigration. All other variables, even sex, may be subject to change and will therefore be either event history variables, variables associated to a specific state, such as benefits or periodic variables.

[1] The word confidential is here used in the same meaning as defined in CR (EC) 831/2002, article 2.

Table 1. A hypothetical event history for a woman immigrating to Norway 22. November 1999. **Bold face** in cell indicates the changed variable.

			Event history variables							
Serial number	Birth year	Dates	Resi-dent	Sex	Marital status	Employ-ment	Resi-dence	Child-ren	Benefits	
123456789	1975	01011992	No							
123456789	1975	22011999	**Yes**	F	U	**Employed**	**Oslo**	0		…
123456789	1975	17062001	Yes	F	U	**Sick**	Oslo	0	86 000	…
123456789	1975	17092001	Yes	F	U	**Employed**	Oslo	0		…
123456789	1975	08022002	Yes	F	U	**Maternity**	Oslo	**1**		…
123456789	1975	25052002	Yes	F	**M**	Maternity	Oslo	1	252 000	…
123456789	1975	08112002	Yes	F	M	**Employed**	Oslo	1		…
123456789	1975	31032003	Yes	F	M	Employed	**Ski**	1		…
123456789	1975	22022005	Yes	F	M	**Unemployed**	Ski	1		…
123456789	1975	27072005	Yes	F	M	Unemployed	Ski	**2**		…
123456789	1975	09042007	Yes	F	M	**Employed**	Ski	2		…
123456789	1975	31052008	Yes	F	**D**	Employed	Ski	2		…
123456789	1975	…	…	…	…	…	…	…		

Just to give an impression of the content, a simplified list of the tables associated to FD-trygd is given in table 2.

Table 2	**Overview of some content in the FD-trygd database**
1	**Demographic tables** Resident status, Sex, Marital status, Family type, No. of children, No. of births, age of youngest child, Place of residence.
2	**Pensions** Dates for start of: Old age retirement, disability, surviving spouse, contractual pension, occupational pension. All monetary benefits associated with these pensions.
3	**Supports** Dates for start of: Preliminary disability, transitional benefit, family income supplement. Child support and support to education for single parents. All monetary benefits associated with these pensions.
4	**Rehabilitation** Medical rehabilitation, sickness benefits, maternity pay, supplementary benefits
5	**Labour Marked** Employment, employment status, industry (ISIC and NACE), expected working hour, unemployed, daily allowance
6	**Education** Not really being a part of FD-Trygd it can be merged. Contains all degrees taken in the Norwegian school system, universities and colleges with dates and associated codes.
7	**Income from assessment** Periodical variables. All variables relating to the tax assessment.

NSD has requested a 20 % sample of all individuals from FD-trygd for anonymisation at their hand based on rules set by Statistics Norway. FD-trygd grows every year. The database is updated once every year with events that have taken place for the residents during the previous calendar year. The updates are copied to the sample for those who have been selected. For newborn and immigrants the sample is updated through Bernoulli sampling. Emigrants and people who die will not be considered residents any more but their histories will remain in the base. FD-trygd now contains histories from about six million people and the sample about 1.2 million. Since the sample is purely register based no one knows who have been selected to the sample. From a confidentiality point of view this is an advantage compared to interview surveys.

As far as we know, the closest parallel to the FD-trygd sample is the Swedish LINDA sample (Longitudinal Individual DAta for Sweden, Edin and Fredirksson, 2000) which goes back to 1960 and now administered by University of Lund (http://www.ed.lu.se/EN/databases/linda.asp). However, LINDA is a longitudinal cross-sectional sample for Sweden based on the Income Register, and not an event history file in the same sense.

3. PRINCIPLES AND DEFINITIONS

In a comment to NSD's request for a 20 percent sample from FD-trygd the legal experts expressed

"… it must be considered how such a sample can be anonymised so that the data are really anonymous. This is a challenge since the actual file drawn directly from FD-trygd, with education and income added, contains a large number of variables with detailed specifications of values. Here, backward identification will be possible even with a 20 % sample."

In this context it is important to have working definitions of the concepts *anonymous* and *identification*. The security handbook of Statistics Norway says that

"… information is anonymous if so many variables or categories have been removed that one cannot with reasonable means *directly or indirectly identify physical or legal persons."*

This definition may be slightly obsolete, e.g. compared to the definition in CR (EC) 831/2002 article 2, and there may be different understandings of the concept "reasonable means", but the definition can be given a sufficiently flexible interpretation to be used as a working definition.

With *backward identification* we understand the same as what is often termed as *identity disclosure* when the formal identifiers have been removed. The variables that can be used for backward identification are called *identifying variables*. Generally we say that a combination of values for such variables will be considered identifying if it meets the following three conditions.

1. *Uniqueness.* There must be a combination of variable values in the dataset on which a person is unique not only in the dataset but in the entire population.

2. *Visibility.* For someone with access to the dataset the values generating a unique combination must be visible on the person to whom they relate.

3. *Verifiability*. It must be possible to verify with a reasonably large probability that the visible combination of values is unique not only in the dataset but in the entire population.

The value of an event history variable is a state process that may consist of a large number of times and new values/states taking effect at each time. Such a combination of times and states will for many persons, perhaps most persons, be unique for at least one event history variable. And even where such histories are not unique for one event variable it will be in combinations with others. Since FD-trygd is a sample from registers covering the entire population it will, at least in principle, be possible for Statistics Norway to determine exactly which combinations are unique in the population and to which persons these combinations belong. Such information could, also in principle, be used for targeted anonymisation of such people. But since the number of such uniques can be very large and anonymisation will have to take place in NSD it is not practical.

The visibility of the variables determines whether or not it should be considered identifying. In the system for anonymisation of FD-trygd the variables have been given a visibility score ν ranging from 0, not visible to 3, very visible. The variable sex is given ν = 3. But since the variable sex divides the population into two approximately equally large groups it is not very identifying. Quite many event variables are very visible, at least their 'present' values. Associated variables such as exact benefits granted are frequently less visible or not at all. The visibility of variable values has changed a great deal with the increasing amount of information available on the internet. This is a process that will continue and affects what should be understood by “reasonable means”.

For a user who has access to a sample only it will not always be possible to verify with certainty if a person is unique in the population even if it is in the sample. But for many combinations of variable values that can be observed it will be natural to consider it very unlikely that there will exist other persons with exactly the same combination. In this way units in the dataset can be *considered* as unique. If the combination fits a persons whose values are visible he or she will be considered identified in the sample.

Putting together data from different tables in FD-trygd is *Data Integration*. On its 57th plenary meeting in 2009 Conderence of European Statisticians (CES) recommended *Principles and guidelines on Confidentiality Aspects of Data integration Undertaken for Statistical or Related Research Purposes, http://live.unece.org/fileadmin/DAM/stats/publications/Confidentiality_aspects_data_integration.pdf*. In addition to having regard to the Norwegian regulations, the rules for researchers’ access to data from FD-trygd seek to respect recommendations in these principles. The document contains an example business case outline expressing the requirements that should be fulfilled in order to do data integration. This business case has been adapted to the actual situation for FD-trygd.

4. LIMITATION OF INFORMATION

To establish an anonymised dataset and limit the disclosure risk it is necessary to restrict the size and detail of the data placed at disposal. It is important to establish clear rules based in measures for information that are no more complicated than that they can be calculated as a number and applied by the institution responsible for using them (NSD). The exact criteria will be formulated when some experience on distribution has been gathered. Until then, deliveries will have to be done with special care. This section describes what such criteria should meet and some of their elements.

The criteria to be established have to meet realistic ‘enemy’ scenarios. This means that we must consider how disclosures are most likely to take place and which incentives there may be for a researcher to break professional secrecy and in some way or another sell or hand over confidential data. The rules must also be able to stand scrutiny from investigating journalists asking critical questions about our confidentiality practice. At the same time the rules must be transparent to the researcher in such a way that they can understand the consequences of the confidentiality methods on their analyses. The rules do not need to meet an intentional intruder with unlimited resources and unlimited willingness to disclose the identity of as many as possible in a dataset. We do not believe such an intruder exists. The most likely scenario as we see it is the accidental recognition of some well known person in a dataset. Such events should be made rare through our confidentiality measures although we will not be able to exclude them completely. For the rest we rely on the secrecy which is part of the researcher’s contract.

To create one complete anonymised 20 % sample with all the variables is out of question. The anonymisation will have to be done separately for every application in a way that meets the specific need of the researcher and at the same time our need to minimise disclosure risks.

When data from the 20 % sample are placed at disposal for researchers this takes place according to the *need to know* principle. This means that the dataset placed at disposal must be restricted in

- sample size
- variable scope
- Length of event history period (years)
- detail for each variable
- Details for dates.

to what the researcher can substantiate that he or she needs for the project. If the variable scope becomes too large one has to consider

- taking different samples with different scope of variables for different analyses.

Different samples should be selected independently. Two such samples may have some common variables. Serial numbers that can identify random overlap should be replaced with a number unique within each sample.

Limitation of variable scope, sample size and period is the first step in the preparation of a sample to researchers. The effect of limiting the sample size will primarily be to limit the number of individuals that can be identified. The limitation of variables, detail for each variable and period serves to limit the possibilities for identifications and how much detail will be disclosed if identification takes place.

The restrictions imply that the researchers must specify with great precision which analyses they want to do and which data they need for each of them. How to balance the ballpoints in a best way both for disclosure protection and for the researcher is a challenge. It must be done in cooperation with the researcher. The idea is that if a measure of risk can be calculated given the values of all the ballpoints, then it will be possible to define a maximum risk that can be accepted for a delivery. With such a maximum risk the researcher applying for a dataset should be able to make trade-off between the ballpoints as long as he or she stays within the risk limitations. This means that the researcher should be able to choose whether he/she wants a large sample size with a small variable scope and little detail or a small sample size with a larger scope and more detail as long as the risk stays within the limit.

Theory for calculating disclosure risk in a cross-sectional dataset based on a given set of variables with given detail exists and has been implemented and can be manipulated in μ-Argus (Franconi and Polettini 2004). An extension of that theory that can give us a practical tool for quantifying disclosure risk for event history data is needed.

Lacking an operational risk measure we have so far decided to set an upper limit on the sample size for each sample delivered. This limit has been set to 10 percent of the target population for the study as it is represented in the 20 % sample. This means about 2 percent of the total target population.

Specific limitations have also been set on a number of variables, in particular the most visible ones. For each variable a *standard representation* has been chosen as the crudest level of detail for the variable that can be given to the researchers if the variable is included in the delivery at all. If a researcher wishes more detail he/she must be able to justify the need. If the need is justified some variables can be delivered with more detail.

All datings are given with the precision of month (YYYYMM) even if many of them exist with the precision of day. This cannot be relaxed.

The demographic variables are very visible (visibility $\nu = 3$) and many of them are therefore given with a reduced standard representation compared to the original data. Since demographic variables are generally important in most research we have attempted to do this in a gentle way. Place of residence is at NUTS 2 level, with seven regions.

A pension state for a person must be considered very visible while exact benefits associated with them are far less so but are considered more private. However, the pension state itself is so central that manipulating their values will render their data almost worthless. Instead the benefits are rounded to approximate values.

Social supports and rehabilitation variables are of more temporary character than pensions and their histories more complicated and slightly less visible. But they are also core variables and difficult to modify without making them useless. Rather, benefits are rounded like for pensions.

Employment states in the labour market are very visible and some associated variables like occupation and industry are very identifying. Occupation is therefore only given with one digit and industry only in 13 categories.

Incomes are given with quintiles of positive values as standard representation.

Education is given with five levels as standard.

5. SUMMARY

The exercise described in this paper breaks a wall in that it attempts an anonymisation of event history data based on registers to researchers. As far as we know this has not been attempted before. The social security system that has produced the data is costly. Learning how it actually works is a condition for improving it and making it more efficient.

NSD is now testing deliveries based on the preliminary rules. In 2012 the rules will be evaluated based on the experience from the deliveries. This will give us more information that will be used to develop the rules to be more flexible. Learning which variables from the database are most demanded will enable us to focus on them in improving the methods. We will also try to establish more specific risk measures that can create more precise frames for the anonymisation. This will enable more flexibility in doing trade-off between sample size and detail and variable scope.

REFERENCES

Franconi, L. and Polinetti, S. (2004). Individual risk estimation in μ-Argus: a review. In: Domingo-Ferrer, J. (Ed.), *Privacy in Statistical Databases.* Lecture Notes in Computer Science pp 262-272. Springer.

UNECE (2009). Confidentiality Aspects of Data integration Undertaken for Statistical or Related Research Purposes. United Nations. Geneva

Edin, P-A. and Fredriksson, P. (2000). LINDA – Longitudinal Individual DAta for Sweden. *Working Paper No. 2000:19*, University of Uppsala, Sweden (http://linda.nek.uu.se).

Andersen, A., Lien, S. and Siverstøl, Ø. (2007). FD-trygd Variabelliste. Notater 2007/17 Statiastics Norway (In Norwegian).