Explorations of multi-level methods and ecological inference techniques in the analysis of “Life courses in Context”

Paper prepared for the International Congress of Historical Sciences, Sydney, 3-9 July 2005

Peter Doorn
Data Archiving & Networked Services (DANS)
peter.doorn@dansdata.nl

Luuk Schreven
Netherlands Institute for Scientific Information Services (NIWI)
Statistics Netherlands (CBS)
luuk.schreven@niwi.knaw.nl or lshn@cbs.nl

1. Introduction

In studies on the history of population in the 19th and 20th centuries, two sources of data are predominant: population censuses and population registers. In many countries, data from these sources are available on the individual level. In the case of the Netherlands, most individual census records have not been preserved. Only for recent years (1960, 1971 and 2001) the anonymized individual records are available for research purposes. What is preserved, however, are about 42,500 pages of published aggregate tables over the period 1795 – 1971 and some 240,000 pages of unpublished detailed aggregate tables for the years 1947 – 1971. These last tables were never published and are only available for referencing at Statistics Netherlands.
For all census years a lot of information is available through these aggregate tables on the municipal level, and since 1849 some basic data are also available at the sub municipal level of the rural settlement or urban living quarter. This information partly overlaps, partly supplements the information from the population registers. In the project “Life Courses in Context”, all aggregate tables from the censuses and a sample of the population registers are being digitized.

Traditionally, historians have put a lot of effort into the development of methods for nominal record linkage. Research on the reconstitution of families and the wish to make longitudinal analyses of individual life cycles formed an important motive for

developing these methods. However, historians have by and large neglected methods to link information across levels of aggregation. This is peculiar, because a variety of multi-level techniques are available and have been widely applied to research problems in the social sciences such as sociology, economics and geography. Moreover, historians generally have to deal with fragmentary information, which is often available at different aggregation levels. This is true for Historical Demography as well as for other historical disciplines.

This paper will explore the potential of two categories of methods and techniques to analyze data across levels of aggregation, using the population registers and censuses from the project "Life Courses in Context". On the one hand, the paper will explore multi-level techniques, which make it possible to incorporate variables into a statistical analysis at a higher aggregated level and to produce good estimates of their effects. In this way aggregated data about, for example, a research subject's place of residence – its population, geographical location and economic circumstances – can be combined with individual variables such as gender and year of birth. On the other hand, it will explore analysis of the opposite type: research at the aggregate level incorporating information about individual people. Again, such "ecological inference" research is still rare in the study of history.

2. Life Courses in Context

Since 1991 the International Institute of Social History (IISG – Internationaal Instituut voor Sociale Geschiedenis) is constructing a database containing micro-level data on the Dutch population. This project is called the Historical Sample of (the population of) the Netherlands (HSN – Historische Steekproef Nederland). Furthermore the Netherlands Institute for Scientific Information Services (NIWI – Nederlands Instituut voor Wetenschappelijke Informatiediensten) has been digitizing published census data since 1997. In 2001 these two institutes of the Royal Netherlands Academy of Arts and Sciences (KNAW – Koninklijke Nederlandse Akademie van Wetenschappen), joined hands in preparing an application for funding from the Netherlands Organisation for Scientific Research (NWO – Nederlandse Organisatie voor Wetenschappelijk Onderzoek). The joint project was dubbed "Life Courses in Context" in which the life courses consist of the HSN and the context is made up by

the digitized census data. The application was granted and funding exceeding 3 million Euro became available in the summer of 2002. On top of that the Royal Academy granted another 600,000 Euro in funds. Other participants in the project are the Historical Databank of Dutch Municipalities (HDNG – Historische Databank Nederlandse Gemeenten) and Statistics Netherlands (CBS – Centraal Bureau voor de Statistiek).

The aim of "Life Courses in Context" is to develop a collaboratory for the study of 19th and 20th century population history. At the core of the project stands the HSN data base with about 40,000 individual life courses of people born in the period of 1863 – 1922 (the HSN database). This database with micro-data will be supplemented with aggregate data from the published Dutch censuses, which were held between 1795 and 1971.

The data sets to be produced by the digitization project will be on different levels of aggregation. The HSN data set is explicitly on the level of the individual and household, and will include personal names. The census data is on a variety of aggregation levels, which will not be the same throughout the period 1795-1971. For the most recent census years, micro-data is available, although the records are anonymous. For most census years, the data are only available in tabular form. For some years, the degree of tabular detail is very high. This is especially the case for the handwritten tables (available since 1930 for the labour census and since 1947 for the population and housing census), but also the published censuses of 1889 and 1899 are very detailed.

Combining these two datasets has some explicit advantages. First of all since a census covers the whole of the population, it can be used as a check on the data collected in the sample when comparing variables available from both datasets. Secondly some variables within the two datasets will be complementary, thus more data will be available to the researchers. When comparing the datasets, the clear advantage of the HSN is that it is a longitudinal dataset (i.e., individual life histories are reconstructed from the sources), whereas the census data are snapshots, these data are transversal (cross-sectional). The advantage of the census data on the other hand is that these data provide much more regional detail then the HSN could ever provide within the confinements of statistical certainty. Lastly, combining the HSN with census data can identify certain individuals from the aggregate tabular census data. For the more recent census years this will result in privacy issues which have to be dealt with.

2.1 Life Courses – Historical Sample of the Netherlands

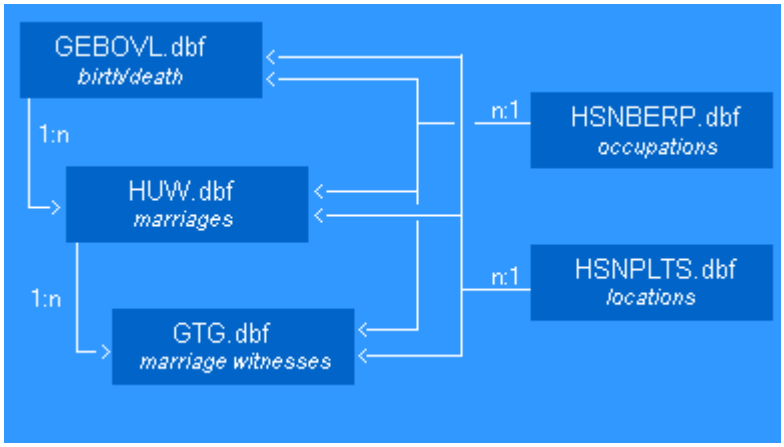
The Historical Sample of the Netherlands (HSN) strives to construct life histories as completely as possible for a representative portion of the nineteenth and twentieth century population in the Netherlands. The sample, for this purpose, has been drawn from all persons born in the Netherlands between 1812 and 1922. Ultimately, the HSN database will include information on an individual level from 77,000 persons on subjects like age at marriage, religious affiliation, the number of children born, occupation, birth place, literacy, social network and migration history.

The dynamic population registration system introduced in the Netherlands in 1850 has resulted in an archive that is unique in international terms. The registers are dynamic in the sense that they do not merely record a situation at a particular moment in time (a snapshot), but that all changes in a subject's address, family size and migration are noted, hereby creating a longitudinal record. The registers are unique because they record details of where inmigrants came from and where outmigrants went to, thus enabling the inclusion of the complete life courses of migrants in the dataset. The HSN project involves the systematic retrieval of the data on life courses from these population registers. Although the population registers were already in use in 1850, the input of life-course data will only take place for subjects born in 1863 or later. There are two reasons for this: first, the registers did not function properly everywhere in the Netherlands during the first few years of their existence. Second, new regulations concerning population registration were introduced in the course of 1862. As a result, the design of the system was modified, and every household was then re-registered. From 1870 onwards the records are actually fairly accurate. Many subjects born after 1870 can also be found in the personal record cards archive at the Central Bureau for Genealogy, so that their migration history can be traced in the reverse direction, thus minimising the risk of "losing" a subject.
The subjects are selected by taking a simple random sample from the birth registers for 1812-1922. The aim is to secure a sample size of 77,000. This is just over half of one per cent of the total number of births, assuming around 14.5 million people were born in the Netherlands in this period. A sample of 77,000 is sufficient for drawing

statistically reliable conclusions on subpopulations of two per cent or more of the population born in the Netherlands during the period.

Besides the life course data from the population registers, data for each individual are systematically collected from the records kept in the public archives. Primarily, these are birth certificates, death certificates, personal record cards and marriage certificates. The birth certificates include information on the person born, as well as the names, addresses, ages and occupations of the parents. The death certificates include the most recent place of residence and most recent occupation of the deceased, and information on his/her spouse(s); certificates of deceased children provide a second indication of the occupational title of the father (the person reporting the death) as well as a double-check on illiteracy. On the basis of these data, it is now already possible to research topics such as childhood mortality and migration patterns for the whole of the Netherlands. Marriage certificates give details of the occupational titles, literacy (signature), and place of residence of the bride and bridegroom, their parents and the witnesses (usually friends or family of the couple). These certificates will enable scholars to research topics such as social and geographical mobility, marital mobility and literacy.

Figure 2.1.1 Structure of the HSN database



At a later stage, information from the population registers, land registers and tax records will be added. These sources are extremely rich, providing information on the family structure, pattern of migration, further occupational history, and the income and wealth of the subject (and sometimes of his or her relatives).

With these characteristics the data set will form a basic resource for research into historical questions concerning problems in the areas of demography, sociology, epidemiology, socio-economics and social geography. The importance of the HSN for the researcher is fourfold. In addition to being an important source for research into social developments in the 19th and 20th centuries and a control database which researchers can use to compare their own research population with, the HSN database also acts as a foundation for the collection of new data and a source of expertise within the field of data collecting. In practice, this is achieved by maintaining a data structure that can be used by individual researchers, and by consistently using the database as a starting point in subsequent research, both by expanding the number of subjects included (oversampling) and enriching the database by introducing supplementary data for specific groups of subjects. For researchers, it cuts both ways. Not only can they use the material already input, they also have access to the software and expertise developed by the HSN. This expertise can be seen as an important by-product of the data-entry work carried out over the past ten years. In return for the use of the software and the data already recorded, the HSN requires researchers to add to the dataset any new data they collect in the course of their research, thus ultimately making it available to other researchers too. This way a data centre will be developed which will function as a centre for quantitative research on life courses.
The nationwide coverage of the dataset ensures that regional variations can be identified, whereas current research involving such topics is usually local in nature and necessarily excludes migrants.

There are comparable projects outside the Netherlands, notably in Québec and Sweden, which in the last twenty years have collected an enormous amount of data. The Dutch data file is distinguished from such foreign examples by the fact that we use a random sample design and that research takes place on a national in stead of a regional level.

2.2 Context – Digitized Dutch census data; 1795 – 1971

National population censuses are one of the fundamental sources of information on conditions in a country. In addition to the population size, the population census

generally contains information on the structural characteristics of a population, such as age, gender, marital status, religion, household status, occupational activity and nationality. Containing this type of data, censuses have always played an important part in historical and social-science research. However since only a limited number of original copies of the 200 or so published volumes of the Dutch censuses have survived and many of these are now in poor conditions, NIWI has been participating in different projects in order to digitize Dutch census data since 1997.

Within the territory of the Netherlands traditional censuses were held between 1795 and 1971. The census of 1795 was conducted by the French occupiers who formed the Batavian Republic. The objective of this census was to register the population of the different parts of the Netherlands in order to construct a new administrative system for the country. When the dust settled after the Napoleonic Wars the Kingdom of the Netherlands was restored and the first official Dutch census was held in 1829 by a Royal Decree. This decree instructed censuses to be held every ten years. In 1879 the content of the Decree was captured in a Census Act.
By 1899 a national statistical agency, Statistics Netherlands (CBS), was created to process the ever growing data collected by the decennial censuses. Of course Statistics Netherlands would also collect and process other data.
Between 1899 to 1930 decennial censuses were conducted by Statistics Netherlands, not only focusing on the population, but also on occupation and housing. The 1940 census was postponed because of the German occupation. After the war, the newly formed Dutch government needed information on the condition of the country as quickly as possible; a population, occupation and housing census was held in 1947 instead of 1950. The connection with the international community was again established by the 1960 census. The last general population census in the Netherlands was held in 1971.
There are three main objectives connected to each of the fourteen official Dutch censuses. The first objective is statistic in its nature; to determine the size of the population on a fixed point in time. The second objective is administrative; to probe and improve the reliability of the Dutch population registers. The third objective is a social one; to examine the demographic and social-economic characteristics of the population. A fourth and final objective emerged on the basis of the more advanced

population, occupation and housing censuses which were held from 1899 onwards: to provided data to facilitate domestic policy making.

Table 1. Overview of censuses held in the Netherlands; 1795 – 1971

Year	*Census*	*Pages*	*Vol.*	*Detailed tables*
1795	*1st integral Dutch population census (Batavian Republic)*	*191*	*2*	
1829	*1st general population census (Royal Decree 1828)*	*18*	*	
1839	*2nd general population census*	*85*	*1*	
1849	*3rd general population census*	*1,165*	*12*	
1859	*4th general population census*	*1,184*	*3*	
1869	*5th general population census*	*889*	*3*	
1879	*6th general population census (Census Act 1879)*	*2,262*	*12*	
1889	*7th general population and occupation census*	*10,223*	*26*	
1899	*8th general population, occupation and housing census (by Statistics Netherlands)*	*9,925*	*27*	
1909	*9th general population, occupation and housing census*	*4,144*	*14*	
1919	*Housing census*	*191*	*1*	
1920	*10th general population and occupation census*	*1,953*	*10*	
1930	*11th general population, occupation and housing census*	*2,353*	*11*	
1947	*12th general population, occupation and housing census*	*1,325*	*12*	*31,000*
1956	*Housing census*	*345*	*3*	*47,000*
*1960***	*13th population census*	*1,809*	*18*	*75,000*
*1971***	*14th general population, occupation and housing census (Census Act 1970)*	*4,503*	*38*	*87,000*
	Total	*42,565*	*193*	*240,000*

* The 1829 census was not separately published. Results were listed in amongst others the 1859 census publications.
** Original micro data available for research purposes.

When preparing for the 1981 census Statistics Netherlands met considerable resistance from the Dutch population. In 1971 2.3 percent of the population had been unwilling to cooperate; by 1981 tests showed that this number had risen to an average of 26 percent for the country as a whole. Within some of the big urban areas this number even amounted to more than 50 percent. The 1981 census was first postponed and later cancelled. To compensate the lack of a general census after 1971, Statistics Netherlands used register counts complemented by random sample surveys in order to meet international obligations (i.e. EuroStat and United Nations regulations). The register counts provided the municipal data on population, housing and other topics. This data was supplemented by big (random sample) surveys in which non-registers information was gathered. From the 1990s onwards efforts were made to gather data annually from different governmental agencies. The 2001 census is referred to as the Virtual Census, it was held by cross linking records from different digital sources to develop census data.

Table 2. "Census" methods after 1971

Year	*Method*
1981	*Register count of Population and housing; survey Labour force and housing needs*
1991	*Register count of Population; survey Labour force and housing needs*
1992 and onwards	*Annual count of housing supply*
1995 and onwards	*Annual count of the population structure*
1998 and onwards	*Annual household statistics*
2001	*Virtual Census*

Abolishing the traditional general census brought some clear advantages as well as some apparent drawbacks. The drawbacks mainly concern the loss of regional detail in the data. Regionally splitting the survey data used in the "censuses" after 1971 beneath the municipal level, will make them less reliable; below the standards employed by Statistics Netherlands. Furthermore it is no longer possible to probe and improve the reliability of the population registers. The advantages of these new "census" methods are the lower costs of the register counts, sample surveys and Virtual Census as compared to a traditional census and the quick results.

Within the census digitization projects the focus has always been on the traditional censuses held between 1795 and 1971. When the Life Courses in Context project is finished NIWI aims to have digitized all aggregate published census tables. These tables will be available through the web and although the data is on an aggregate level, the degree of detail within the tables is very high. Even though the Life Courses in Context project is scheduled to run until July 2006, NIWI is proud to inform that most of the tables are already available through the internet in the form of Excel spreadsheets. Also some basic documentation is downloadable from the web as well as the digitized source material; in the first of NIWI's census digitization projects all published census volumes (more than 40,000 pages) were scanned and published on five CD-roms. Furthermore, whenever possible, the data will be harmonised over time using harmonisation schemes developed by the iPUMS-international project. This enables international comparison as well as temporal comparison. While most data is already available on the website the status of the data is not always final. Some validation and correction of data-entry still has to be done.

The digital census database that ensues from this project will be an important resource for historical and social-science research. National historical census projects have been, or are being, carried out in a number of countries, including the US, the UK, Ireland, France, Norway, Denmark, Germany, Russia and Austria. Some of these projects are based on the original source material, which means databases can be constructed at individual level. In the case of the Dutch censuses individual data has only survived for the most recent census years; 1960, 1971 and 2001. This data is being used to create 1 percent samples for the iPUMS-international project.

2.3 Comparison of Life Courses (HSN) and Context (census) variables

Before exploring the different multi-level methods and ecological inference techniques in the analysis of "Life courses in Context" data it is valuable to review the differences between the two data sources.

Table 2.3.1 Comparison between variables within Life Courses in Context data sources

HSN micro data (birth, death and marriage registers)	***Census aggregate data***
Date (and hour) of birth	*Age (groups)*
Place of birth (municipality of birth certificate)	*Municipality of birth (nationality, ethnicity)*
Sex	*Sex*
Date of marriage	*NA*
Place of marriage	*NA*
Marital status	*Marital status*
Occupational title	*Occupation (-al group, sector)*
NA	*Religion*
Address	*Neighborhood/municipality*
NA	*Characteristics of dwelling (housing censuses)*
Age of parents at birth	*NA*
Signature (proxy for illiteracy)	*Educational attainment*
Relationships to family members and witnesses	*Position in Household*
Date (and hour) of death	*NA*

As mentioned above there are some distinct differences between the data sources that will have to be clear to every researcher using the data. First of all census data cover a complete population at a certain point in time, whereas the data collected by the HSN is a longitudinal sample of the population. Second, the level of the data differs;

whereas the HSN provides individual data, the census data is on an aggregate level. A third difference between the two sources is the variables within the data sets. Combining these two datasets, both with their own strengths and weaknesses, can create some very interesting research opportunities that were not possible before. The remainder of this paper will be aimed at the exploration of analyses techniques combining data from both sources.

3. Combining data across levels of aggregation

The sources that historians work with are often limited in their information value: they are usually incomplete, vague and/or partly incorrect with respect to the problem we are trying to solve. This is one reason why historians often use different sources in combination with each other. In quantitative analysis, especially in population studies, a lot of effort has been put in the accurate linking of information on individuals, and a variety of methods has been designed for what is known in the literature as nominal record linkage.[1] Record linkage techniques have always concentrated on the linking of individuals who occur in different sources (for instance, persons in birth, marriage and death registers). It is however remarkable that historians have rarely tried to combine data from sources of unequal level of aggregation, to investigate how information on groups could supplement information on individuals or vice versa. In geography and (other) social sciences such as economics, sociology and politicology, the combination of data across levels has received much more attention. In this section we look at different approaches to combine data from different levels of aggregation, taking the micro-data from the population registers on the individual level and the aggregate tables from the censuses of the Netherlands in the 19th and 20th centuries as an example.

What does the linking or combined study of individual data from the HSN with the aggregate tables of the censuses have to offer? The linkage of individual life-course data with cross-sections from censuses will be enriching in various ways. For a start, the population censuses may offer a context for the individual-level and family-level

[1] 'Record linkage', ed. S. W. Baskerville, P. Hudson and R. J. Morris, *History and Computing*, special issue, iv (1992). G. Bloothoot, 'Assessment of Systems for Nominal Retrieval and Historical Record Linkage', *Computers and the Humanities*, 1998, vol. 32, no. 1, pp. 39-56.

data. The combination of the different sources will create new opportunities for analysis in the following ways:
First, the HSN is a sample, whereas the censuses cover the whole population. The combined study of both sources makes it possible to validate the quality of the sample to the whole population for corresponding variables in the years of the censuses. This may result in weights to correct for possible biases in the sample.
Second, the variables in the HSN and the censuses are partly complementary. It can be fruitful to use ecological variables from the censuses to explain individual changes in the life cycle.
Third, the combination is also interesting from a methodological point of view: the mere fact that the combined analysis of individual and aggregate sources is rare in historical research, makes it useful to test the potential.

In the following sections, we will outline three approaches to combine the data across levels of aggregation:
1. Aggregating individual data
2. Multi-level or cross-level analysis
3. Disaggregating aggregate data

3.1. Aggregating individual data

As mentioned above, the census information on individuals in the population has not been systematically preserved before 1960. The individual data have been aggregated in many ways, both spatially and otherwise (e.g. in municipalities, occupational groups, age groups, religious denominations, etc.). This aggregate data has been published in the form of cross-tables. The cross-tables are often multi-dimensional and contain many categories, as well in the rows as in the columns of the tables

The Historical Sample of the Netherlands (HSN) should be random and unbiased, and the censuses may offer means to check whether this is true. Aggregating the HSN to variables and categories as they occur in the census tables is the most straightforward way to combine the two sources, which of course will mean that the detail of the individual (or family) level will be lost. This amounts to aggregating the HSN data for

cross-sections at census years. However, the complexity of the longitudinal HSN files (on births, marriages and deaths), makes it not an easy task to select the correct persons in order to create the comparable variables or categories for the dates on which the censuses were held. Moreover, the HSN sample is too small to make groups of sufficient size for small units. Another complication is that in the course of time, various groups and regions in the HSN have been over sampled, which means that comparison with the censuses will show deviations that have been "built in" by the sample design. Finally, of course, the censuses themselves will be far from perfect, meaning that when statistical deviations are found, it is not clear whether they are caused by the census or by the population registers.

Figure 3.1.1: example of a multi-dimensional table from the Dutch census

Plaatselijke indeeling.	Doel der gestichten.	Woningen			Bewoonde		
		bewoond.	leegstaand.	in aanbouw.	schepen		Wagens
					varende	als woonhuis dienend	
					waarvan de bewoners werkelijke woonplaats in de gemeente hebben.		
1	2	3	4	5	6	7	8
1. Aagtekerke		116	5				
2. Verspreide huizen		59	2				
Totaal Aagtekerke (Z.)		175	7				
1. Aalsmeer (Streekdorp) met de onderdeelen Oosteinde en Uiterweg: Tehuis voor ouden van dagen	V	1					
1. Aalsmeer (Streekdorp) met de onderdeelen Oosteinde en Uiterweg: Carmelieten klooster	Kl						
1. Aalsmeer (Streekdorp) met de onderdeelen Oosteinde en Uiterweg: Andere woningen, schepen, enz.		1583	21	44	4	14	
1. Kudelstaart		142	2	1	1	11	
1. Kalslagen		26					
2. In den Schinkelpolder		25				1	
2. In den Oosteinderpoelpolder		109	1	2			
2. In den Stommeerpolder		18					
2. In den Hornmeerpolder		12					
2. In den Zuiderlegmeerpolder		21					
3. Overige verspreide huizen		14					
Totaal Aalsmeer (NH.)		1951	24	47	5	26	

3.2. Multi-level analysis

At least since Durkheim's famous book on suicide (originally published in 1897), social scientists have paid attention to effects of the surroundings on the individual. From the 1950s and 1960s onward terms such as 'contextual' (Lazarsfeld 1959), 'neigborhood' (Cox 1972), 'structural' (Blau 1960) and 'compositional' (Davis et al. 1961) effects are used. In the more recent literature since the 1980s, the terms 'multilevel' and 'cross-level' effects are used to indicate the variety of effects that operate on more than one level. Another term that is relevant in this context is the 'ecological fallacy'. In a well-know article Robinson (1950) indicated that relations on the individual level can deviate from correlations on an aggregate level. The ecological fallacy is a widely recognised error in the interpretation of statistical data, whereby inferences about the nature of individuals are based solely upon aggregate statistics collected for the group to which those individuals belong. This fallacy assumes that all members of a group exhibit characteristics of the group at large. Stereotypes are one form of ecological fallacy.

Here the issue at stake is how the individual, longitudinal data from the population registers and the cross-sectional, mostly aggregate data can enrich each other. The censuses offer many background variables that are not available in the population registers of the HSN. On the other hand, the detail of individual persons and households of the HSN is not available for most census years. In analyses at the individual level, ecological effects of higher levels (groups, municipalities, regions) may be taken in consideration.

The family of multi-level or cross-level analysis techniques offers solutions for variables operating at different levels that may influence each other (Snijders & Bosker 1999; Goldstein 1994; Hox 1995; Gerland 1996). The combined analysis of the data sets of the HSN and censuses will stimulate the application and further development of such techniques in historical research.

It should be noted that the actual linkage of records is not the aim here. In multilevel analysis the central objective is to statistically explain a phenomenon, in which effects of higher levels of scale (ecological or context variables) are included in the analysis

(for example: school results can be influenced by characteristics of the individual, the family, the school and the wider environment (neighbourhood, city, etc.).[2]

3.3. Reconstruction of individual records from aggregate tables

The approach described in this section is the reverse of that described in section 3.1. However, disaggregating aggregate data is much more complex than the opposite. As mentioned before, the census tables are often very detailed. There are many cells in the tables which are empty and there are numerous cells with the value of 1 (person). Only in the published census tables of 1971 values have been rounded to 0 or multiples of 5 to prevent the identification of individuals for privacy reasons. But in many older tables we can indicate individuals, although we do not know who those individuals are by name.

There is also no obvious linkage possible of individuals in different tables, although we know a variety of characteristics about them. Among several tables, a certain degree of overlap in variables or categories of information exists. In nominal record-linkage methods, the names of individuals are usually the starting point of the linking, but these are not present in the Dutch censuses. By combining overlapping variables in different tables (such as occupation, date of birth, address) the most probable linkages can be established.

Statistical Disclosure Control and synthetic estimation methods -

In research into Statistical Disclosure Control (SDC) it is attempted to prevent that individuals can be recognized or identified in tables of a low degree of aggregation: *'Statistical Disclosure Control (SDC) comes into play when data about individual entities such as persons, households, businesses etc. are released by a data disseminator such as a statistical office. Before releasing data to the outside world, the statistical office has to make sure, within a reasonably short time, that no*

[2] Goldstein, Harvey, 1994, 1999R, *Multilevel Statistical Models*. London, Edward Arnold.
Griffin, Mark A. & David A. Hofmann, 1997, "Hierarchical Linear Models in Organizational Research: Cross-level Interactions." 1997 *Research Methods Forum* No. 2 (Summer). http://www.aom.pace.edu/rmd/1997_forum_hlm_models.html
Hox, J.J. , 1995, *Applied multilevel analysis*. Amsterdam, TT-Publicaties.
Snijders, Tom & Roel Bosker, 1999, *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*. Sage, Beverly Hills, etc.

individual entity can be recognised from these data. The identification of individual entities can be avoided by modifying the data in appropriate ways'.[3] In the SDC-project that was carried out in the context of the EU Esprit-programme, universities and statistical offices from the Netherlands, Italy and the United Kingdom put the optimisation of data-modification central. Both methodological research was conducted and SDC-software was developed (Argus).

Synthetic estimation methods deal with attempts to reconstruct synthetic individual records from aggregate data. This technique is related to the estimation of the filling of a table given the marginal totals. For example, Paass (1989) describes a stochastic modification algorithm used to construct a synthetic sample X from different input sources, the sources being independent samples or summary statistics from an underlying population. The first step in the process is to construct an X as a best fit to the data by a maximum likelihood or minimum cost criterion, and the second step is to generate a sample with a cost value near the minimum which also has maximum entropy. Rubin (1993) proposes that only "synthetic data" rather than actual micro-data should be released as a method of disclosure risk avoidance. The synthetic data would be generated using multiple imputations. They would look like individual reported data and would have the same multivariate statistical properties.[4] Similar techniques may be developed to attempt to reconstruct synthetic individual records from detailed census tables, e.g. for those of 1889, 1899 and 1947.

Ecological inference - Historians have not systematically investigated the possibility to reconstruct individual records from aggregate tables, although some isolated examples exist. Jan-Bernd Lohmöller and Hartmut Bömermann applied the

---

[3] Bethlehem, J.G., Keller, W.J., and Pannekoek, J., 1990. "Disclosure Control of Microdata", *Journal of the American Statistical Association*, 85, no. 409, pp.38-45.
Elliot, Mark, 1996, "Attacks on census confidentiality using the Sample of Anonymised Records: an analysis" (Paper for presentation to 3rd International Seminar on Statistical Confidentiality, Bled, Slovenia, 2nd October 1996).
Gerland, Patrick, 1996, Socio-economic data and GIS: datasets, databases, indicators and data integration issues. Paper presented at the UNEP/CGIAR (Consultative Group on International Agricultural Research), Arendal III Workshop on Use of GIS in Agricultural Research Management. Norway, June 17-21, 1996.
Mokken, R.J., Kooiman, P., Pannekoek, J., and Willenborg, L.C.R.J., 1992. "Disclosure Risks for Microdata," *Statistica Neerlandica*, Vol. 46 no. 1, pp. 49-67.

[4] Rubin, D., 1993, "Discussion, Statistical Disclosure Limitation," *Journal of Official Statistics*, Vol. 9, No. 2, pp. 461-468.

Goodman approach to ecological inference techniques in their research on "voter movements to Nazism" in an article published in 1992.[5]

Fairly recently, a lot of attention has been paid to reconstruct individual behaviour from aggregate data. Probably the most comprehensive contribution to the development of techniques to do so is made by Gary King.[6] According to King and colleagues, ecological inference is the best and often the only hope of making progress where individual-level data are absent. “Ecological inference is the process of extracting clues about individual behaviour from information reported at the group or aggregate level”. Especially after King’s book published in 1997, an explosion of statistical research into the problem of ecological inference has appeared. In the last five years we have seen numerous new models, innovative methods and novel computation schemes. Unfortunately, ecological inference is an especially difficult case of statistical inference. So far, it remains a challenge to apply the technique to the Dutch censuses. In appendix 1 a rather simplistic example of a possible disaggregation of a hypothetical census table is given.

4. Conclusion and directions for future research

This paper made a plea for more interest by historians for the linkage of data from different levels of aggregation: micro data and ecological data. The data from the project Life Courses in Context offers an excellent opportunity to do so, but there are plenty of similar data sets in other countries that offer opportunities for such multi-level approaches.

Of the different approaches, aggregating micro data on life histories to levels that exist in published census tables, is the simplest. Multi-level analysis does not actually create data links, but offers analytical tools to include context variables in the explanation of individual behaviour. Ecological inference methods are technically complex, but are very promising to distil individual information from aggregate tables. The next step is to elaborate on the approaches described in this paper empirically. All

[5] J.B. Lohmöller, & H. Bömermann, ‘Kontingenztafelschätzung aus Aggregatdaten’, in: *Historical Social Research* 64:17 (1992) No. 4, p. 3-69.

[6] G. King, *A solution to the ecological inference problem: reconstructing individual behavior from aggregate data.* (Princeton U.P., 1997). Gary King, Ori Rosen & Martin A. Tanner (eds.), *Ecological Inference: New Methodological Strategies* (Cambridge U.P., 2004), Series: Analytical Methods for Social Research.

ingredients are available: the data and the techniques are there. What we are missing is a researcher who wants to take on the challenge to carry it out. Anybody who is interested to do this work is welcome to contact us.

References

Agrawal, Rakesh & Ramakrishnan Srikant, 2000. *Privacy preserving data-mining.* IBM Almaden Research Center (San Jose, CA). http://www.almaden.ibm.com/cs/people/srikant/papers/sigmod00.pdf

Bethlehem, J.G., Keller, W.J., and Pannekoek, J., 1990. "Disclosure Control of Microdata, " *Journal of the American Statistical Association*, 85, no. 409, pp.38-45.

Brink, T. van den, 'The Netherlands population registers', in: *Sociologia Neerlandica* 3, p. 51-63

Doorn, P.K., 'Modern monnikenwerk: digitalisering van historische bronnen', in: *Groniek* 151 (march 2001), p. 211-226.

Elliot, Mark, 1996, 'Attacks on census confidentiality using the Sample of Anonymised Records: an analysis' (Paper for presentation to 3rd International Seminar on Statistical Confidentiality, Bled, Slovenia, 2nd October 1996).

Gerland, Patrick, 1996, Socio-economic data and GIS: datasets, databases, indicators and data integration issues. Paper presented at the UNEP/CGIAR (Consultative Group on International Agricultural Research), Arendal III Workshop on Use of GIS in Agricultural Research Management. Norway, June 17-21, 1996.

Goldstein, Harvey, 1994, 1999R, *Multilevel Statistical Models*. London, Edward Arnold.

Erwich, B. & J.G.S.J. van Maarseveen (ed.), *Een eeuw statistieken. Historisch-methodologische schetsen van de Nederlandse officiële statistieken in de twintigste eeuw* (Voorburg/Amsterdam 1999)

Gordon, C., *The Bevolkingsregisters and their use in analyzing co-residential behaviour of the elderly* (NIDI report no. 9: Den Haag 1989)

Griffin, Mark A. & David A. Hofmann, 1997, 'Hierarchical Linear Models in Organizational Research: Cross-level Interactions.' 1997 Research Methods Forum No. 2 (Summer). http://www.aom.pace.edu/rmd/1997_forum_hlm_models.html

Hox, J.J. , 1995, *Applied multilevel analysis*. Amsterdam, TT-Publicaties.

Knotter, A. & A.C. Meijer (ed.), *De gemeentelijke bevolkingregisters, 1850-1920* (Den Haag 1995)

Maarseveen, J.G.S.J. van & P.K. Doorn (ed.), *Nederland een eeuw geleden geteld: een terugblik op de samenleving rond 1900* (Amsterdam 2001)

Maarseveen, J.G.S.J. van (ed.), *Algemene tellingen in de twintigste eeuw* (Voorburg/Heerlen 2002)

Mandemakers, K., 'The Netherlands: Historical Sample of the Netherlands', in: P.K. Hall, R. McCaa & G. Thorvaldsen (ed.), *Handbook of international historical microdata for population research* (Minneapolis 2000)

Mokken, R.J., Kooiman, P., Pannekoek, J., and Willenborg, L.C.R.J., 1992. "Disclosure Risks for Microdata," *Statistica Neerlandica*, Vol. 46 no. 1, pp. 49-67.

Paass, G., 1989, "Stochastic Generation of a Synthetic Sample from Marginal Information," Proceedings of the Bureau of the Census Fifth Annual Research Conference, Bureau of the Census, Washington, DC, pp. 431-445.

Rubin, D., 1993, "Discussion, Statistical Disclosure Limitation," *Journal of Official Statistics*, Vol. 9, No. 2, pp. 461-468.
Schulte Nordholt, E., M. Hartgers & R. Gircour (ed.), *The Dutch Virtual Census of 2001. Analysis and Methodology* (Voorburg/Heerlen 2004)
Snijders, Tom & Roel Bosker, 1999, *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*. Sage, Beverly Hills, etc.
Vos, S. de, *De omgeving telt: compositionele effecten in de sociale geografie*. PhD Dissertation, University of Amsterdam (1997).
Vulsma, R.F., *Burgerlijke stand en bevolkingsregister* (Den Haag 2002)

Internet sources:
www.lifecoursesincontext.nl
www.volkstelling.nl
www.iisg.nl/~hsn
www.cbs.nl
www.ipums.org

Appendix 1. Example of a reconstruction of individual records from two hypothetical cross tables

Table 1	*Column variable 1*			
	Column A	*Column B*	*Column C*	*Total*
Row 1	*1*	*2*	*3*	*6*
Row 2	*4*	*5*	*1*	*10*
Row 3	*2*	*3*	*0*	*5*
Total	*7*	*10*	*4*	*21*

Table 2	*Column variable 2*			
	Column D	*Column E*	*Column F*	*Total*
Row 1	*0*	*1*	*5*	*6*
Row 2	*2*	*3*	*5*	*10*
Row 3	*1*	*2*	*2*	*5*
Total	*3*	*6*	*12*	*21*

Tables 1 and 2 both contain information on 21 individuals. The two column variables 1 and 2 differ, but the row variable (in the case of the census often the municipality) is identical. Both aggregate tables can be easily transformed into tables with persons as observations (records).

Individualized Table 1		
Person	*Row*	*Col.Var. 1*
1	*1*	*A*
2	*1*	*B*
3	*1*	*B*
4	*1*	*C*
5	*1*	*C*
6	*1*	*C*
7	*2*	*A*
8	*2*	*A*
9	*2*	*A*
10	*2*	*A*
11	*2*	*B*
12	*2*	*B*
13	*2*	*B*
14	*2*	*B*
15	*2*	*B*
16	*2*	*C*
17	*3*	*A*
18	*3*	*A*
19	*3*	*B*
20	*3*	*B*
21	*3*	*B*

Individualized Table 2		
Person	*Row*	*Col.Var. 2*
1	*1*	*E*
2	*1*	*F*
3	*1*	*F*
4	*1*	*F*
5	*1*	*F*
6	*1*	*F*
7	*2*	*D*
8	*2*	*D*
9	*2*	*E*
10	*2*	*E*
11	*2*	*E*
12	*2*	*F*
13	*2*	*F*
14	*2*	*F*
15	*2*	*F*
16	*2*	*F*
17	*3*	*D*
18	*3*	*E*
19	*3*	*E*
20	*3*	*F*
21	*3*	*F*

What is the most likely linkage of the persons from tables 1 and 2? The characteristics of the column variables do not overlap in this example. From every row (e.g. municipality) from the original tables the number of persons per category in the column variable is know. We can use these numbers as marginal totals in the next three contingency tables, in which column variable 1 and 2 in every row variable (municipality) is crossed. The contents of the cells of the table are not known, but we can calculate expected frequencies on the basis of the marginal totals under the hypothesis that both variables are not associated. If an association does exist (or is

hypothesized), this can be included in the calculation of the calculation of the cell frequencies. In this example we assume there is no association.

Row 1				
Column	*D*	*E*	*F*	*Total*
A	*0*	*0*	*1*	*1*
B	*0*	*0*	*2*	*2*
C	*0*	*1*	*3*	*3*
Total	*0*	*1*	*5*	*6*

Row 2				
Column	*D*	*E*	*F*	*Total*
A	*1*	*1*	*2*	*4*
B	*1*	*2*	*3*	*5*
C	*0*	*0*	*1*	*1*
Total	*2*	*3*	*5*	*10*

Row 3				
Column	*D*	*E*	*F*	*Total*
A	*0*	*1*	*1*	*2*
B	*1*	*1*	*1*	*3*
C	*0*	*0*	*0*	*0*
Total	*1*	*2*	*2*	*5*

In the example the expected frequencies are rounded to whole numbers, for example the value 1 in the crossing of columns A en D in Row 2 is reached by multiplying the outer values (2 and 4) and dividing by the outer total (10). In practice, there will often be partly overlapping information in column variables. On the one hand, this leads to redundancy in the individualized tables, that can simply be removed. On the other hand this overlap can improve the linkages. Logically impossible or unlikely combinations (e.g. position in household: child and age: older than 65; marital status: married and age: under 15 years old) can be avoided by building in constraints in the estimation.
With the help of the expected frequencies of the occurrence of combined characteristics per row variable (municipality) the most probable linkages between individuals can now be made:

Person	***Row***	***Col.Var 1***	***Col.Var 2***
1	*1*	*A*	*F*
2	*1*	*B*	*F*
3	*1*	*B*	*F*
4	*1*	*C*	*E*
5	*1*	*C*	*F*
6	*1*	*C*	*F*
7	*2*	*A*	*D*
8	*2*	*A*	*E*
9	*2*	*A*	*F*
10	*2*	*A*	*F*
11	*2*	*B*	*D*
12	*2*	*B*	*E*
13	*2*	*B*	*E*
14	*2*	*B*	*F*
15	*2*	*B*	*F*
16	*2*	*C*	*F*
17	*3*	*A*	*E*
18	*3*	*A*	*F*
19	*3*	*B*	*D*
20	*3*	*B*	*E*
21	*3*	*B*	*F*