Preliminary version

Combination of register and survey data: The Norwegian Generations and Gender Survey

Helge Brunborg and Trude Lappegård
Statistics Norway

Paper prepared for the Seminar on Registers in Statistics - methodology and quality,
Helsinki, 21-23 May 2007

Introduction

The Nordic countries are in the vanguard of social and demographic development in Europe and the world. This can be seen in a number of areas, such as the social welfare system and in demographic areas such consensual unions, gender equality, participation of women in work and society, and participation of men in child rearing and family life. It is, therefore, important to include the Nordic countries in comparative studies of the social and economic differentials and trends.

The Generations and Gender Programme (GGP) is an ambitious and broad attempt to obtain new data about the changing social and demographic landscape in Europe. GGP is a system of national Generations and Gender Surveys (GGS) and contextual databases, which aims at improving the knowledge base for policy-making in UNECE countries. The design for the survey is face-to-face interviews with a large number of questions based on a standard questionnaire.

For Norwegian participation in this programme it would be necessary to use alternative data collection methods, because the costs of the original GGS design would be prohibitive for a high-cost country like Norway. In addition to reducing the costs, alternative data collecting methods have been introduced in order to save interviewing time. The design for the Norwegian Generation and Gender Survey is, therefore, based on data collected from both primary (survey) and secondary (register) sources. The primary data sources are computer assisted telephone interviews (CATI) and self-administered questionnaires on paper.

The Generations and Gender Programme welcomed our design and our use of administrative registers, although with some scepticism because our methodology would be significantly different from the methodology of other countries participating in the programme. On the other hand, our approach was considered to be very interesting and valuable for the development of survey methodology. The Nordic countries are in the front with regard to secondary data sources, with our rich and rapidly developing administrative registers, which we rely more and more on in statistics and analysis.

The Norwegian GGS is covering approximately 460 questions about social and demographic development, mostly questions that cannot be covered by our registers. Fully 80 of these questions, or 18 per cent, will be replaced by information collected from secondary data sources, i.e. administrative registers. The aim of this paper is to outline some issues due to possible instrument effects of combining primary and secondary data sources. There is a need to do methodological work to look at the quality, reliability, validity and comparability of data drawn from primary and secondary sources.

The Generation and Gender Programme

The Generations and Gender Programme (GGP) is a system of national Generations and Gender Surveys (GGS) and contextual databases. The GGS is a panel survey of a nationally representative sample of 18-79 year-old resident populations in each participating country, with at least three panel waves and an interval of three years between each wave. The contextual databases are designed to complement micro-level survey data with macro-level

information on policies and aggregate indicators (Source: http://www.unece.org/pau/ggp/Welcome.html).

The main substantive goal of the programme is to improve understanding of demographic and social development and of the factors that influence these developments, with a particular attention towards relationships between children and parents (generations) and relationships between partners (gender). The GGP sets out to explain demographic choices in forming and dissolving partnerships and having children. The analytic domains covered by the survey include economic aspects of life, such as economic activity, income, and economic well-being, education, values and attitudes, intergenerational relationships, gender relationships, household composition and housing, residential mobility, social networks and private transfers, public transfers, health, and reproductive health. By including all these topics, the GGP covers the important societal aspects of demographic choices in contemporary developed societies.

A questionnaire for GGS has been developed by an international group of leading scholars (http://www.unece.org/pau/ggp/materials/GGS_Qre_Core.pdf). This is intended to be used as much as possible, but have been adjusted in all participating countries to a smaller or larger extent to local conditions or to reduce the interviewing time.

About 15 countries have joined the GGP, including France, Germany, Belgium, Hungary, Italy, Russia, Bulgaria, Czech Republic, Romania, Slovakia, Japan and Australia. Norway is so far the only participating Nordic country, but Finland is investigating the possibilities for joining, with funding as the largest obstacle. Belgium, which also has comprehensive administrative registers, is starting the interviews in 2007. The Netherlands, whose registers are very similar to the Nordic registers, is not formally participating in GGP, but their Kinship and Family Survey is covering many of the same areas as the GGS.

The Norwegian Generations and Gender Programme

Two surveys in one

The Norwegian Generations Survey is called LOGG (Life course, Generations and Gender). It is a combination of two surveys, as it has been expanded to include the second wave of norLAG, a panel survey that was conducted for the first time in 2002-2003.

The questionnaire is based on the GGS questionnaire and the norLAG questionnaire, but had to be shortened to allow for a telephone interview of not more than 45 minutes (30 minutes of GGP questions and 15 minutes of norLAG questions).

Data sources

In addition to the interviews, data are collected through a self-administered postal questionnaire and data from administrative registers. The interviewing started in January 2007 and is expected to finish in the first half of 2008.

Sample

- Sample: National + NorLAG areas (30 municipalities/boroughs)
- Age range: 18-79 years
- Sample size: 26 000 (6 000 interviewed for the second time)

- Letter sent to all sampled respondents before the interview

Register data

Register data are added both before and after the interview:

Before the interviewing begins the PIN (Personal Identification Number), name, address and relationship for the following persons are extracted from the Central Population Register (CPR):

- Respondent
- Spouse (living in the same household or not)
- Children, including those who do not live with the respondent
- Parents (for respondents born after 1952)
- Siblings (living in the same household or not)
- Grandchildren (for respondents born after 1935)
- Cohabitants (from the Dwelling Register)

Several of the survey questions will be replaced with administrative records (also for family members). *After* the survey the following register data are added for each respondent, and also for family members, for:

- Births
- Marital history
- Migration history
- Parental leave: duration and income compensation
- Income and wealth
- Educational activity and attainment
- Social insurance: disability, old age, social welfare, cash for care

A preliminary list of variables to be collected from registers is included (in Norwegian) in Appendix 1.

We are also planning to construct some new variables from register information, for example, the distance between the dwellings of those family members who are identified in the CPR. This will be done by using a programme that estimates distance in km and meters between addresses in Norway. This programme has also been developed to estimate the travelling time by car in minutes between addresses, allowing for speed limits and ferry time when necessary. It may, however, be unrealistic to travel by car from Oslo to Tromsø to visit a child, for example. Because of this we may also utilize a program developed by the Institute for Transport Economics (TØI) that estimates the mode of transportation, travelling time and the costs between all municipalities in Norway, based on observations of travel patterns.

Use of administrative records in surveys

Requirements for use

In order to combine survey and register data certain conditions need to be met. The minimum requirements for this are:

- That there is a unique identifier for the total resident population, which is usually a number called the Personal Identification Number (PIN).
- That the PIN needs to be included in survey for each respondent.

- There need to be available data sources with PIN for the total population with additional variables, especially a CPR (Central Population Register). Censuses are also possible additional data sources.

Reasons for use

When Norway was encouraged to join the Generations and Gender Programme it was envisaged that the most important reason for including interview data in addition to survey data was to save interview time and costs. It is clear, however, that registers are used in combination with surveys for a number of reasons:

1. To save interviewing time and costs.
2. To draw the sample and find the addresses of the respondents.
3. To keep track of respondents for the next waves in panel surveys (panel maintenance).
4. To analyse characteristics of non-respondents.
5. To obtain more detailed data, e.g. on education.
6. To obtain life history data for periods before and after interview, e.g. marital and educational histories.
7. To get data of better quality, e.g. of date of birth, and income and wealth.

Thus, the savings in costs is not the most important reason for combing register and survey data.

Limitations of use

There are, however, also a number of drawbacks or limitations in combining registers with an interview survey:

1. Registers do not include any data on *subjective* variables such as preferences and intentions, such as time spent with children, intended number of children, or feeling of loneliness.
2. Some administrative data are of questionable *validity*, such as disability.
3. Some administrative data are of questionable *reliability*, e.g. residential address and age at leaving the parental home since people do not always live on the address that is registered by the CPR.
4. The *definitions* of certain variables in registers often differ from the intended survey definition, e.g. duration of partnership.
5. Register data values are sometimes *different* from the values obtained in an interview, e.g. income and wealth.
6. Registers usually include *limited or no data* for older cohorts, especially those born before the CPR was established (1964 in Norway). For this reason we decided, e.g., to ask about grandchildren of respondents born before 1952 since the PIN are included for almost all residents born after 1952.
7. Events *abroad* are generally not recorded in registers, such as educational attainment.
8. It is some times more *complicated and costly* to extract data from registers than to collect them in an interview, e.g. occupation and hours of work.
9. Registers are continuously developed and improved and some registers have still not become good enough to be used to replace interview data, such as consensual unions, which utilize the recently introduced Dwelling Register.

10. There is usually a time lag between an interview and the availability of the timeliest administrative variable, e.g. on income in the current year, which in Norway becomes available about 18 months after the end of year in which the income was earned. Similarly, educational attainment in Norway can only be obtained in October 1 in the year after the final exam was completed.
11. Even though some registers are continuously updated, such as the central population register, there is in practice a time lag between extracting information from the register and the day of the interview. This implies that in some cases the household composition of the respondent has changed. For example, the respondent may have moved (or died). This makes the telephone interview more complicated, if some of the information available for the interviewer was collected before the interview and is shown on the screen during the interview session. It is, however, to make the corrections later, after the interview.

Although GGS and LOGG are intended to cover all aspects of the life cycle, there are some variables that we hesitate to use, even though they may relatively cost-effectively be drawn from registers. The reasons for this may be legal, ethical or practical. In Norway, and we believe in other "register" countries as well, the selected respondents receive a letter from Statistics Norway some time before the interview. The letter presents the survey, stresses the importance of participation, and says that information about the respondent from certain registers will be linked to the data collected in the interview. The names of registers need to be included, such as the Central Population Register and the Income register but the letter does not need to mention every variable that will be drawn from the registers. If the respondent is not objecting it is assumed that he or she tacitly agrees to the use of register data for him or her.

For theses reasons, the mentioning of registers such as register the Cause of Death Register and Register of Criminal Cases in the letter to the respondents, would probably scare off many and reduce the response rate. Thus, one reason for not including data from all available registers is that it would reduce the quality of the survey.

In LOGG the situation is complicated further because register data are collected not only for the respondents but also for other persons with a relationship to the respondent, including spouses, cohabitants, children and parents, grandchildren and grandparents. Again, it is assumed that non-refusal can be interpreted as acceptance.

Summary

In this paper we have outlined some issues due to possible instrument effects of combining primary and secondary sources, using the Norwegian participation in the Generations and Gender Programme as an example. New technologies are being used to achieve high response rates and quality combined with low costs. There is a need for more knowledge of possible instrument effects of the combination of such alternative data collection in order to improve and facilitate data collection.

Appendix 1: Preliminary list of variables to be collected from registers

Registeropplysninger om IO (intervju-person) og ektefelle/registrert partner/ samboer for der personnummeret er kjent

- Fødselsår
- Kjønn
- Alder ved intervjuet
- Sivilstatus på trekktidspunktet
- Sivilstatus fra 1967-2007
- Bydel (i Oslo)
- Bostedskommune på trekktidspunktet
- Bostedskommune ved intervjuet
- Bostedskommune 1967-2007
- Tettbygd/sprettbygd
- Sentralitet
- Landsdel
- Fødeland
- Landbakgrunn
- Innvandringskategori
- Statsborgerskap
- Utdanningsnivå 1970-2007
- Utdanningens fagfelt 1970-2007
- Studiekode 1970-2007
- Inntekt (yrke/kapital/samlet) 2006-2007
- Pensjon 2006-2007
- Skatteplikte overføringer 2006-2007
- Barnetrygd 2006-2007
- Bostøtte 2006-2007
- Stipend 2006-2007
- Stønad (grunn/hjelpe) 2006-2007
- Forsørgerfradrag 2006-2007
- Formue 2006-2007
- Kapital 2006-2007
- Gjeld 2006-2007
- Ektefelletillegg/fradrag 2006-2007
- Fødselspenger 1992-2007
- Pensjonspoeng 1992-2007
- Alderspensjon 1992-2007
- Uførepensjon 1992-2007
- Etterlatte pensjon 1992-2007
- AFP (Avtalefestet pensjon) 1992-2007
- Sykepenger (siste hendelse)
- Rehabilitering/attføring 1992-2007
- Overgangsstønad 1992-2007
- Omsorgspoeng 1992-2007
- Kontantstøtte 1998-2007

Registeropplysninger om husholdsmedlemmer utover partner

- Alder
- Kjønn
- Relasjon til IO
- Landbakgrunn
- Fødeland
- Statsborgerskap

Registeropplysninger om IOs barn, barnebarn, mor, far, søsken

- Alder
- Kjønn
- Sivilstatus
- Bostedskommune
- Statuskode (bosatt, død, emigrert)
- Statusdata (dato når ovennevnte inntraff)

Pluss om IOs barn

- Utdanningsnivå 1970-2007
- Utdanningens fagfelt 1970-2007
- Studiekode 1970-2007
- Reisetid (hvis ikke bor sammen)
- Avstand i meter (hvis ikke bor sammen)

Pluss om IOs mor og far

- Utdanningsnivå 1970-2007
- Utdanningens fagfelt 1970-2007
- Reisetid (hvis ikke bor sammen)
- Avstand i meter (hvis ikke bor sammen)