European Population Conference 2006

Topic 9 (Data and Methods)

The IPUMS-Europe project: Integrating the Region´s Census Microdata

Dr. Albert Esteve (Centre d'Estudis Demogràfics)

Prof. Robert McCaa (Univeristy of Minnesota and Minnesota Population Center)

Prof. Anna Cabré (Universitat Autònoma de Barcelona and Centre d'Estudis Demogràfics)

Abstract.- Census microdata are an invaluable resource for social science and policy research. Other sources—such as demographic and labor force surveys—often offer greater subject coverage and detail than do census data, but no alternate source offers comparable sample density, chronological depth, and geographic coverage. This paper describes the IPUMS-Europe project, a consortium lead by the Minnesota Population Center and the Centre d'Estudis Demogràfics to anonymize, harmonize and distribute census microdata of eighteen European countries from the 1960s to the present. The database will contain anonymized microdata samples, encompassing as many as 50 censuses and totaling more than 70 million person records. Custom-tailored extracts will be delivered, at no charge, to bona fide researchers via the Internet. The new database will allow social scientists to make comparisons across European nations during decades of marked demographic change and extraordinary political and economic restructuring.

Introduction.- Census microdata are an invaluable resource for social science research. Other sources—such as demographic and labor force surveys—often offer greater subject focus and detail than do census data, but no alternate source offers comparable sample density, chronological depth, and geographic coverage. A vast quantity of census microdata covering Europe in the period since the 1960s survives in machine-readable form. For much of Europe, census microdata are either unavailable or restricted, and are therefore seldom used. In the United States and Canada, however, census microdata have been available to researchers for almost forty years and have become an indispensable component of social science infrastructure.

Thanks to the support of official statistical agencies in 14 European countries and major funding by the National Institutes of Health and the European Commission Sixth Framework Programme, the Integrated European Census Microdata database, one of the world's largest integrated research infrastructure for the study of human populations, is now under construction. The database will contain anonymized microdata samples encompassing as many as 50 censuses and totaling more than 70 million person records. The National Institutes of Health (NIH) have awarded the Minnesota Population Center (MPC) a major grant to undertake a five-year initiative to create integrated and fully documented samples of over sixty European censuses and micro-censuses from the 1960s to the present (IPUMS-Europe project). The project will integrate and disseminate the census microdata of Austria, Belarus, Bulgaria, the Czech Republic, France, Germany, Greece, Hungary, the Netherlands, Portugal, Romania, Slovenia, Spain and the United Kingdom. In addition, the Centre d'Estudis Demogràfics (CED) has been successful in attracting European Union Sixth Framework Program's support for coordination, dissemination and harmonization. EU funds have already provided for an inaugural workshop, held in Barcelona in July in 2005, at which census experts discussed harmonization strategies to integrate European census microdata across space and time (Coordinating the Integration of European Census Microdata, CIECM project). The Sixth Framework Program will also support a three year initiative to build and European web-based dissemination extract site, housed at the Centre d'Estudis Demogràfics, which will make the European microdata and metadata more widely available for scholarly and educational research (Disseminating the Integrated European Census Microdata, DIECM project). Finally, to fully capitalize on the potential of European census microdata, a third project has been approved to design harmonizations for ~50 priority variables for each census & country for which microdata samples are entrusted to the project ensuring that coding schemes that reflects census practices of European states as well as the principles and recommendations of Eurostat with regard to census concepts and nomenclatures (Harmonizing the Integrated European Census Microdata, HIECM project).

Confidentiality protections.- The IPUMS-Europe project distributes integrated microdata of individuals and households only by agreement of the corresponding national statistical offices and under the strictest of confidence. These protections involve three elements:

1. dissemination agreements between the University of Minnesota and each NSI
2. user licenses between, on the one hand, the University of Minnesota or other authorized distributor such as the CED, and, on the other, each researcher
3. data protection measures to prevent the identification of individuals, families or other entities in the data.

Data Quality.- In addition to providing harmonized codes for variables and accompanying documentation, the IPUMS-Europe project is carrying out a variety of additional tasks to improve data quality, not all of which have been implemented in the first release of the data. These tasks include the following:

- Cleaning data to eliminate duplicate records, inappropriately merged households, and other errors
- Developing internal consistency checks to maximize data integrity. This includes, for example, examining consistency between age and marital status, occupation, and

school attendance; looking for persons with multiple spouses for countries in which this is not an accepted custom; and checking for agreement between household and individual characteristics.

- Implementing allocation procedures to impute values for missing or inconsistent data items, using logical edits together with probabilistic "hot deck" methodology. A data quality flag identifies allocated data items.

Creating constructed variables to simplify data analysis, including family interrelationship variables. A system of logical rules identifies the record number within each household of the individual's mother, father, or spouse, if they were present in the household. These pointers allow users to automatically attach the characteristics of these kin or to construct measures of fertility and family composition. Other constructed variables describe family and household characteristics at the individual and household level (such as family and subfamily membership, family and subfamily size, and number of own children).

Harmonization.- European census samples employ differing numeric classification systems and reconciliation of these codes is a major effort. Variables must be easy to use for comparisons across time and space. This requires that we provide the lowest common denominator of detail that is fully comparable. On the other hand, we must retain all meaningful detail in each sample, even when it is unique to a single dataset.

For most variables, it is impossible to construct a single uniform classification without losing information. Some samples provide far more detail than others, so the lowest common denominator of all samples inevitably loses important information. Composite coding schemes offer a solution. Similar to that used by the International Labor Organization for occupations and industries, we apply composite coding to each variable to retain all original detail, and at the same time provide comparable codes across countries and censuses. The first one or two digits of the code provide information available across all samples. The next one or two digits provide additional information available in a broad subset of samples. Finally, trailing digits provide detail only rarely available.

IPUMS-Europe users will have access to at least 50 priority variables harmonized according to intra-European coding schemes and disseminated by the DIECM project. These variables cover all census topics (Demography, Education, Economic Activity, Migration, Household Composition, and Dwelling characteristics) and are available in the vast majority of countries (See Figure 1).

Figure 1. Selected variable topic availability, by country - 2000 Census round

	AUS	BLR	BUL	CZ	FRA	GER	GRE	HUN	POL	POR	ROM	RUS	SLV	SPA	UK
PERSON VARIABLES															
Demographic and social															
Relationship to household head	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Age	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Sex	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Maritial Status	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Age at first marriage	.	.	X	X	.	X	X	X	X	.	X	.	.	.	.
Citizenship	X	X	X	X	X	X	X	X	X	X	.	X	X	.	.
Religion	X	.	X	X	.	.	.	X	.	X	X	.	X	.	X
Language	X	.	X	X	.	.	.	X	X	.	X	X	X	X	X
National and/or ethnic group	.	.	X	X	.	.	.	X	X	.	X	.	X	.	X
Children ever born	X	X	X	.	.	X	X	X	X	.	X	X	X	.	.
Education															
Literacy	X	X	X	.	.	.	X	.	.	X	.	X	.	X	.
School attendance	.	X	X	.	X	X	X	X	X	X	X	X	X	X	X
Educational attainment	X	.	X	X	X	X	X	X	X	X	X	X	X	X	X
Economics															
Employment status	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Time worked	X	.	.	.	X	X	X	X	.	X	X	.	.	X	X
Unemployment duration	.	.	.	.	X	X	.	.	X	.	X	.	.	.	X
Occupation	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Industry	.	.	.	.	X	X	X	.	.	.	X	X	X	X	X
Socio-economic status in employment	X	X	.	X	.	.	X	.	X	.	.	.	X	.	X
Class of worker	X	X	.	.	X	X	X	X	.	.	X	X	X	X	X
Place of work	X	X	X	X	X	X	X	X	.	.	X	X	X	X	X
Mode of transport to work	X	.	X	X	X	X	.	X	.	X	.	.	X	X	X
Length and frequency of journey to work	X	.	.	X	.	X	X	X	.	.	.	.	X	X	X
Disability	.	.	X	.	.	.	.	X	X	X	.	.	.	.	X
Migration															
Place of usual residence	X	X	X	.	X	X	X	X	X	X	X	.	X	X	X
Size of place, urban/rural	X	X	X	.	X	X	X	X	X	X	X	.	X	.	.
Place of birth, within country	X	.	X	.	X	.	X	.	X	X	X	X	X	X	X
Place of previous residence	.	X	X	X	X	X	X	X	X	X	X	.	X	X	X
Country of birth	X	X	X	.	X	.	X	.	X	.	X	X	X	X	X
Reason for immigration	.	X	.	.	.	.	X	.	X	.	.	X	X	.	.
Country of citizenship	.	X	X	.	X	X	X	X	.	.	X	X	.	.	.
Year/period of immigration	.	.	.	.	X	X	X	.	.	.	.	.	X	.	.
HOUSEHOLD VARIABLES															
Household characteristics															
Location	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Tenure status	X	X	X	.	X	.	X	X	.	X	X	X	X	.	X
Rent	X	X	.	.	.	.	.	.	.	X	X	X	.	.	.
Number of vehicles	.	.	.	X	X	.	.	.	.	.	.	.	.	X	X
Living quarters															
Ownership	X	X	X	X	X	.	X	X	X	X	X	.	X	X	X
Vacancy status	.	.	.	X	X	.	X	X	X	X	X	.	X	.	.
Number of occupants	.	X	X	.	.	.	X	.	.	X	.	X	X	.	.
Number of rooms	X	X	X	.	X	.	X	X	X	X	.	X	X	X	X
Useful and/or living floor space	X	.	.	X	X	.	.	X	X	.	X	X	X	X	.
Facilities															
Electricity	.	X	X	.	.	.	X	.	.	X	X	X	X	.	.
Water / hot water	X	X	X	X	.	.	X	X	X	X	X	X	X	.	.
Sewage	.	.	X	X	X	.	X	X	X	X	X	X	X	.	.
Toilet	X	X	X	X	X	.	X	X	X	X	X	.	X	.	.
Bathing facilities	X	X	X	X	X	.	X	.	X	X	X	X	X	.	X
Type of heating	X	X	X	X	X	.	X	X	X	X	X	X	X	X	X
Piped gas	.	X	.	X	.	.	.	X	X	.	X	X	X	X	.
Building characteristics															
Type with regard to constructuion / use	X	X	X	X	X	.	X	X	X	X	X	X	X	.	.
Period of construction	.	X	X	X	X	.	X	X	X	.	X	X	X	X	.
Position of dwelling in the building	.	.	.	X	X	X	X	.	.	.	X	X	X	.	.
Number of dwellings in the building	X	.	X	X	X	.	X	.	.	.	X	.	X	.	.
Construction materials	.	X	X	X	.	.	.	X	.	.	.	X	X	.	.

Note: a single variable topic in this table can represent multiple variables in the source data.

AUS: Austria, BLR: Belarus, BUL: Bulgaria, CZ: Czech Rep., FRA: France, GER: Germany, GRE: Greece, HUN: Hungary, POL: Poland, POR: Portugal, ROM: Romania, RUS: Russia, SLV: Slovenia, SPA: Spain, UK: United Kingdom

Potential Impact.- The availability of consistent microdata for all of Europe over a broad time span will have a profound effect on the practice of social science research. The new database will allow social scientists to make comparisons across European nations during decades of marked demographic change and extraordinary political and economic restructuring, including the shift to free market economies in Eastern Europe and the growth and development of the European Union. In concert with data from other census integration projects, these European data will also stimulate international comparative research across continental boundaries. The data will result in an outpouring of new scientific and policy-relevant research on population aging, economic transformation, demographic change, international migration, and many other topics. The European microdata series will help policymakers and scholars make informed decisions about the most obvious topics of analysis.