UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

STATISTICAL OFFICE OF THE EUROPEAN UNION

Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses
(The Hague, The Netherlands, 10-11 May 2010)

Working paper 19
4 May 2010

Topic (vii) of the provisional agenda

REGISTERS: THE MAIN SUPPLIER OF TOMORROWS' RESEARCH WAREHOUSES?

Register-based vs. traditional census samples: The IPUMS-International experience[*]

Note by the Minnesota Population Center, Minneapolis, MN USA and Centre d'Estudis Demogràfics, Autonomous University of Barcelona

I. SUMMARY

1. Register-based censuses are planned for 17 UNECE member states in 2011, of which eight participate in the IPUMS census microdata initiative to disseminate samples to researchers on a restricted access basis. The Netherlands is the only country currently entrusting register-based samples to the IPUMS project. Constructing samples from register-based data poses few serious technical issues, but there are at least two significant obstacles: administrative (laws and regulations regarding confidentiality and dissemination), and methodological (sample unit and variable availability). This paper addresses the second issue. An analysis of IPUMS user statistics suggests that high-precision, household samples, rich in demographic, social and economic detail, are essential if the microdata are to be used widely. Usage statistics for 47,623 IPUMS extracts are analyzed to reveal the variables in greatest demand. Probably the most striking finding is that three IPUMS constructed variables rank among the top ten percent of variables requested: spouse's location in household, mother's location, and father's location. These variables can only be constructed from household samples. Register based censuses may make it possible to offer a greater range of variables than conventional censuses. However, if households cannot be constructed from the registers, then the data will suffer from a lack of demand among IPUMS-International users. High precision, richly detailed household samples are essential for high quality

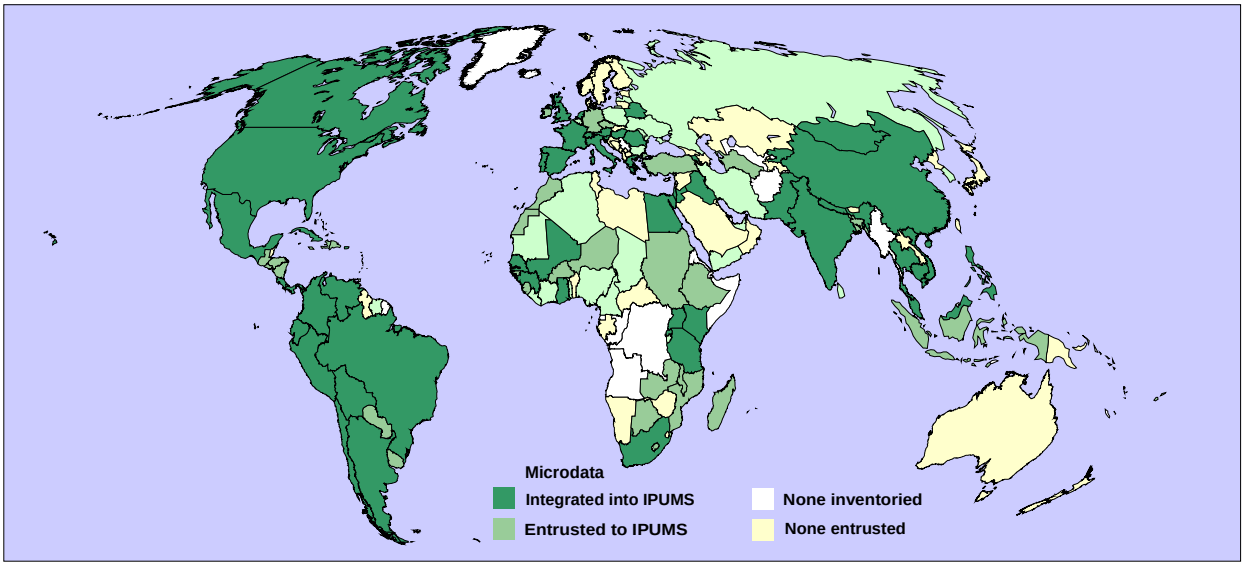
[*] Research for this paper was funded in part by the National Institutes of Health of the United States, grant HD047283 European and Asian census microdata harmonization project (IPUMS-EurAsia) and Harmonizing Integrated European Census Microdata (HIECM), funded by the European Union, Research Infrastructures Action, FP6-026033.

research based on census microdata—regardless of whether the census is conventional or register-based.

II. IPUMS-International: a massive, widely used, global resource for restricted access census microdata

2. IPUMS-International (www.ipums.org/international) archives, integrates, and disseminates high precision, richly detailed microdata from national population and housing censuses. This massive data infrastructure, totalling more than 320 million anonymized, integrated person records representing almost 90 million households, encompasses 55 countries and 158 censuses. Thanks to sustained funding by the National Science Foundation and the National Institutes of Health (USA) as well as exceedingly generous cooperation from National Statistical Offices worldwide, the database is expanding at the rate of 5-10 additional countries per year (see Table 1). The project is led by the University of Minnesota Population Center in partnership—for the censuses of Europe—with the Centre d'Estudis Demogràfics, Autonomous University of Barcelona (www.iecm-project.org).

Table 1. Status of census microdata for countries participating
in the IPUMS-International collaboratory, March 2010:
3 shades of green--Integrated (dark), Integrating (medium), and Negotiating (light)



3. Twenty European countries participate in the IPUMS initiative (number of sets of microdata contributed in parentheses; * = microdata are integrated and being disseminated): *Armenia (1), *Austria (4), *Belarus (1), Bulgaria (0), the Czech Republic (2), *France (7), Germany (8—includes GDR, FRG and microcensuses), *Greece (4), *Hungary (4), Ireland (8), *Italy (1), the *Netherlands (3), *Portugal (3), *Romania (3), *Slovenia (1), *Spain (3), *Switzerland (4), Turkey (0), Ukraine (0), and the *United Kingdom (2—to be expanded to 6). Integration of microdata through the 2000 round of censuses is complete for fourteen European countries. The 2011 IPUMS launch is scheduled to incorporate samples for three European countries—France (2006 census), Ireland and Germany—as well as a half dozen non-European nations (Cambodia, Egypt, Jamaica, Indonesia, Malawi, Morocco, and Nicaragua). Additional launches are planned for successive years, integrating 2010 round census samples as expeditiously as they become available. Of the 20

European countries currently participating in IPUMS, the following are planning register based censuses for the 2010 round: Austria, Czech Republic, Germany, the Netherlands, Slovenia, Spain, Switzerland, and Turkey.[1] In the 2000 round only the Netherlands conducted a register based census. As we shall see, a sample for the 2001 census of the Netherlands was constructed and integrated into the IPUMS database, but it is among the least used, with fewer than 200 extracts.

4. Although access to the IPUMS-International microdata is free of cost, usage is restricted to bona-fide researchers who agree to abide by stringent conditions of use. IPUMS disseminates extracts, custom-tailored to the precise research needs of each user. The average IPUMS extract consists of a mere 10 variables. This contrasts with the practices of most statistical offices where census microdata are disseminated as complete sets, consisting of a data dictionary and an entire sample. Typically, under this modus operandi, when requests are fulfilled, all researchers receive the same set of data and documentation. Given the massive size of the IPUMS-International database, disseminating the full set of variables and unvarying size of samples is impractical. Instead, with IPUMS, the researcher requests an extract from the database, in which selections are made for:

- country (or countries)
- census year(s)
- variables (age, sex, educational attainment, etc.)
- sub-populations (e.g., female heads of households aged less than twenty five years along with all other co-resident persons in the selected household)
- and sample density (either as a percent or number of cases).

The IPUMS extract engine fulfils the request by generating a dataset containing only the requested microdata and the corresponding codebook (available in 4 flavors: generic, SPSS, SAS or STATA). Additional comprehensive metadata are available from the web-site, both as documents and in interactive form.

5. At the UN-ECE Expert Group Meeting on Statistical Data Confidentiality, November 2005, we explained the IPUMS-International data dissemination procedure as follows[2]:

> When the extract is ready (usually in a matter of minutes), the researcher is notified by email that the data should be retrieved within 72 hours. A link is provided to a password-protected site for downloading the specific extract. The data are encrypted during transmission using 128-bit SSL (Secure Sockets Layer) encryption standard, matching the level used by the banking and other industries where security and confidentiality are essential. The researcher may then securely download the file, decompress it and proceed with the analysis using the supplied integrated metadata consisting of variable names and labels.

This method of dissemination has weathered the test of time, and indeed as usage soars, the rapid acceleration of internet transmission speeds has validated this approach.

---

[1] Conference of European Statisticians, "Main Results of the UNECE-UNSD Survey of the 2010 Round of Population and Housing Censuses," United Nations Commission for Europe, October 2009, pp. 2-3. http://www.unece.org/stats/documents/2009.10.census.htm

[2] Robert McCaa and Albert Esteve, "*IPUMS-Europe: Confidentiality measures for licensing and disseminating restricted access census microdata extracts to academic users*," Monographs of official statistics: Work session on statistical data confidentiality. Luxembourg: Office for Official Publications of the European Communities, p. 40.

III. IPUMS-International Usage Statistics (through January 2010).

6. 47,623 extracts have been downloaded from the IPUMS-International site, averaging over 1,000 per country for the 44 countries represented in the database (Table 2). Nonetheless usage varies greatly by country, from as few as 69 for the 1999 census of the Kyrgyz Republic to a high of 6,706 for the six censuses of Mexico. The top ranked sample is Mexico 2000, with 2,221 extracts. The sole register based census sample, Netherlands 2001, ranks in the bottom quartile, with only 193 extracts. As we shall demonstrate below, this is not due to the fact that it is register-based but rather because it is a sample of persons, not households, and offers only 25 variables in contrast to the more heavily used household samples with 90 or more variables.

7. Mexico, Brazil and Colombia predominate in usage not only because they offer lots of variables and a long series of household samples covering a half century of population history, but also due to the fact that many Latin American emigrants reside in the United States or Spain and thus it is possible to analyze these populations in a single integrated database, whether they reside in the country of birth or in the two most important countries of emigration. In addition, all the Latin American samples, as well as those for the United States and Spain, are high precision and richly detailed with extensive information on migration, as well as economic, social and demographic variables of households as well as individuals.

Table 2. Extract Rank and Sample Details for the Top Five and all European Countries

Rank	Country	Sample %*	Variables (n)*	Years of census samples	Extracts
1	Mexico	10	120	1960p, 70, 80, 95, 2000, 05	6,706
2	Brazil	5	106	1960, 70, 80, 91, 2000	4,497
3	United States	5	92	1960, 70, 80, 90, 2000, 05	3,898
4	Colombia	10	120	1964p, 72, 85, 93, 2005	2,785
5	France	5	99	1962, 68, 75, 82, 90, 99	2,194
13	Greece	10	89	1971, 81, 91, 2001	1,221
14	Spain	5	99	1981, 91, 2001	1,201
19	Hungary	5	74	1970, 80, 90, 2001	874
21	Portugal	5	96	1981, 91, 2001	809
22	Romania	10	97	1976, 92, 2002	772
23	Austria	10	75	1971, 81, 91, 2001	755
28	United Kingdom	3	47	1991, 2001p	508
30	Netherlands	1	33	1960p, 71p, 2001p	428
32	Belarus	10	84	1999	267
40	Italy	5	81	2001	127
43	Slovenia	10	80	2002	78
Total extracts from the IPUMS-International database for 44 countries (130 samples) Feb. 1, 2010					47,623

*2000 round census; refers to all integrated variables, including IPUMS constructed variables.
"p" = person sample; all other samples are of households

8. One can surmise from statistics in Table 2 that register based samples should be richly detailed—with at least 70 variables—high-precision household samples—at least five percent—if

they are to be widely used by IPUMS-International researchers. For the seven countries converting from conventional to register-based censuses in the 2010 round and participating in the IPUMS initiative, we hope that the microdata samples will be highly compatible with conventional samples. The usage statistics indicate that a series of high precision household samples with at least 70 variables in each will be extensively used.

The only register based sample currently in the IPUMS database is the 2001 census of the Netherlands. Unfortunately, all the Dutch samples, including the historical datasets for 1972 and 1960, are low precision person samples with the fewest variables of any in the IPUMS-International database (see Table 3 and Appendix A).

Table 3. Household Variables in IPUMS-International Samples: France and the Netherlands Compared

		Variable label	France		The Netherlands			IPUMS
Household variables			2006	1999	1960	1971	2001	Constructed
1	PERSONS	Person records in the household	X	X	X	X	X	X
2	WTHH	Household weight	X	X	X	X	X	
3	GQ	Group quarters status	X	X	X	X	X	
4	UNREL	Number of unrelated persons	X	X	.	.	.	X
5	REGIONW	Continent and region of country	X	X	X	X	X	
6	REGNFR	Region, France	X	X	.	.	.	
7	ENUTS1	NUTS1 Region, Europe	X	X	.	.	.	
8	ENUTS2	NUTS2 Region, Europe	X	X	.	.	.	
9	OWNRSHP	Ownership of dwelling	X	X	.	.	.	
10	FUELH	Fuel for heating	X	X	.	.	.	
11	AUTOS	Automobiles available	X	X	.	.	.	
12	ROOMS	Number of rooms	X	X	.	.	.	
13	BATH	Bathing facilities	X	X	.	.	.	
14	HHTYPE	Household classification	X	X	.	.	.	X
15	NFAMS	Number of families in household	X	X	.	.	.	X
16	NCOUPLS	No. married couples in HH	X	X	.	.	.	X
17	NMOTHRS	Number of mothers in household	X	X	.	.	.	X
Total Household and Dwelling Variables			36	17	4	4	4	6
Total Person Variables			57	49	17	18	21	17

9. Statistics Netherlands, as one of the founding members of the IPUMS initiative and the only statistical institute inclined to entrust a register based sample to the project, was unable to take advantage of the experience of other statistical offices. With the first sets of microdata entrusted to IPUMS, little consideration was given to the set of variables offered. Recently, due to the substantial amount of work required to integrate census microdata and metadata and the widespread

cooperation from many statistical offices, the IPUMS-International collaborator has adopted the following minimum standards for microdata integration and dissemination[3]:

1) household samples (person samples only where no other microdata exist such as for the 1970 round and earlier censuses)
2) high precision—5% minimum, 10% preferred
3) broad set of variables—typically all that are available, omitting only those required to protect statistical confidentiality.
4) detailed codes—for age (single year to age 85), occupation (3 digit ISCO), country of birth (detailed), etc.—suppressing only those codes required to protect statistical confidentiality, such as those with low frequencies.

10. It is not surprising that usage of the Dutch samples is limited because they contain relatively little information. The four household variables are purely technical and are at a distinct disadvantage when compared with the 17 variables in the sample of the 1999 census of France. With respect to persons, the Dutch samples offer barely a score of variables, compared with 49 for the 1999 French sample. It is hoped that it will be possible to construct new samples for the Netherlands—both the conventional as well as the register-based censuses, including the upcoming census of 2011. If so, the samples should include an ample array of dwelling, household and person variables, and, the sampling unit should be the household rather than the person.

11. 3,065 researchers have qualified for access to the IPUMS-International database, representing 81 countries. The average user made fifteen extract requests. In total more than one-half million variables (n=503,184) have been downloaded. The fact that the average extract consists of a mere ten variables shows that IPUMS-International users are researchers, not hoarders. Their extract requests are parsimonious, limited to specifically what is needed to address well-defined research questions.

12. Of the more than 700 integrated variables available to researchers, 32 of the most commonly extracted are listed in Table 4. The top 8 encompass 4 demographic variables (marital status, relationship to head, age and sex), 2 economic (employment status and class of worker), and one each social (educational attainment) and technical (person weight).

[3] Robert McCaa, Wendy Thomas, Albert Esteve and Antonio López Gay, "Entrusting census microdata and metadata for timely integration and dissemination via the IPUMS-EurAsia and IECM initiatives, 2010-2014", Joint UNECE/Eurostat Meeting on Population and Housing Censuses (28-30 October 2009, Geneva, Switzerland)

Table 4. IPUMS-International integrated variables by number of extracts

Rank	Label	Extracts	Mnemonic	Comment
1	Educational attainment	16,910	EDATTAN	
2	Employment status	16,397	EMPSTAT	
3	Marital status	15,952	MARST	
4	Person weight	14,915	WTPER	Technical variable
5	Relationship to head	13,769	RELATE	
6	Age (single years to 85+)	13,145	AGE	Grouped age n=3,169
7	Sex	12,766	SEX	
8	Class of work	11,294	CLASSWK	
9	School attendance	7,182	SCHOOL	
10	Occupation ISCO recode	7,166	OCCISCO	
11	Ownership of dwelling	6,998	OWNRSHP	
12	Years of schooling	6,680	YRSCHL	
13	Literate	6,435	LIT	
14	Urban/rural	6,399	URBAN	
15	Industry-general code	6,289	INDGEN	
16	Household weight	5,935	WTHH	Technical variable
17	Children ever born	5,772	CHBORN	
18	Nativity (native/foreign born)	5,723	NATIVTY	
19	Occupation	5,682	OCC	
20	Country of birth	5,479	BPLCTRY	
21	Religion	5,343	RELIG	
22	Industry	5,160	IND	
23	Location of spouse in household	4,494	SPLOC	IPUMS constructed
24	Rule for locating spouse	3,790	SPRULE	IPUMS constructed
25	Number of children surviving	3,717	CHSURV	
26	Place of residence 5 years ago	3,652	MGRATE5	
27	Location of mother in household	3,559	MOMLOC	IPUMS constructed
28	Total household income	3,522	INCTOT	Household variable
29	Location of father in household	3,447	POPLOC	IPUMS constructed
30	Earned income	3,312	INCEARN	
31	Waste disposal	3,105	SEWAGE	
32	Consensual union	3,067	CONSENS	

13. One of the more surprising rankings is the presence of four IPUMS constructed variables among the top 30 requested: location in household of spouse, mother, and father and the rule used in locating the spouse—what experienced IPUMS users fondly refer to as the "LOC" variables. Researchers use these variables to study the joint characteristics of spouses and characteristics of parents relative to their children. The variables are constructed by inference from the relationship to head variable, age, sex, marital status, order of individuals listed in the household, and a few other variables. The heavy use of the "LOC" variables indicates their great importance for analyzing individuals in relation to their spouses, mothers and fathers.

IV. Conclusion.

14. As the practice of register based censuses spreads to many countries, it is important to take into account that researchers require more than simply a couple of dozen variables about individuals. As evidenced by the IPUMS-International usage statistics, many researchers seek to study individuals in their household contexts. If register-based censuses provide microdata solely about individuals, it will be impossible to use such data for a substantial fraction of the social science research agenda. Where registers permit the construction of households, and statistical offices are inclined to facilitate access to household samples with a rich array of variables, the IPUMS-International usage statistics show that researchers will certainly make ample use of them.

Appendix A. Availability of Variables: 5 samples compared

		Variable label	France		The Netherlands			IPUMS
Household variables			2006	1999	1960	1971	2001	Constructed
1	PERSONS	No. of person records in household	X	X	X	X	X	X
2	WTHH	Household weight	X	X	X	X	X	
3	GQ	Group quarters status	X	X	X	X	X	
4	UNREL	Number of unrelated persons	X	X	.	.	.	X
5	REGIONW	Continent and region of country	X	X	X	X	X	
6	REGNFR	Region, France	X	X	.	.	.	
7	ENUTS1	NUTS1 Region, Europe	X	X	.	.	.	
8	ENUTS2	NUTS2 Region, Europe	X	X	.	.	.	
9	OWNRSHP	Ownership of dwelling	X	X	.	.	.	
10	FUELH	Fuel for heating	X	X	.	.	.	
11	AUTOS	Automobiles available	X	X	.	.	.	
12	ROOMS	Number of rooms	X	X	.	.	.	
13	BATH	Bathing facilities	X	X	.	.	.	
14	HHTYPE	Household classification	X	X	.	.	.	X
15	NFAMS	Number of families in household	X	X	.	.	.	X
16	NCOUPLS	No.of married couples in household	X	X	.	.	.	X
17	NMOTHRS	No. of mothers in household	X	X	.	.	.	X
Household/dwelling variables in 2006 sample not in 1999			19					
		Sum of Household variables	36	17	4	4	4	6
Person Variables								
1	PERNUM	Person number	X	X	X	X	X	X
2	WTPER	Person weight	X	X	X	X	X	X
3	MOMLOC	Mother's location in household	X	X	.	.	.	X
4	POPLOC	Father's location in household	X	X	.	.	.	X
5	SPLOC	Spouse's location in household	X	X	.	.	.	X
6	PARRULE	Rule for linking parent	X	X	.	.	.	X
7	SPRULE	Rule for linking spouse	X	X	.	.	.	X
8	STEPMOM	Probable stepmother	X	X	.	.	.	X
9	STEPPOP	Probable stepfather	X	X	.	.	.	X
10	POLYMAL	Man with more than one wife linked	X	X	.	.	.	X
11	POLY2ND	Woman is second or higher order wife	X	X	.	.	.	X
12	FAMUNIT	Family unit membership	X	X	.	.	.	X
13	FAMSIZE	No of own family members in HH	X	X	.	.	.	X

14	NCHILD	Number of own children in household	X	X	.	.	.	X
15	NCHLT5	No. of own children under age 5 in HH	X	X	.	.	.	X
16	ELDCH	Age of eldest own child in household	X	X	.	.	.	X
17	YNGCH	Age of youngest own child in HH.	X	X	.	.	.	X
18	RELATE	Relationship to household head	X	X	X	X	X	
19	ERELATE	Relationship to head, Europe	X	X	X	X	X	
20	AGE	Age	X	X	X	X	X	
21	AGE2	Age, grouped into intervals	X	X	X	X	X	
22	SEX	Sex	X	X	X	X	X	
23	MARST	Marital status	X	X	X	X	X	
24	EMARST	Marital status, Europe	X	X	X	X	X	
25	NATIVTY	Nativity status	X	X	X	X	X	
26	BPLFR	Region of birth, France	X	X	.	.	.	
27	EBPLNT1	Region of birth, Europe, NUTS1	X	X	.	.	.	
28	EBPLNT2	Region of birth, Europe, NUTS2	X	X	.	.	.	
29	CITIZEN	Citizenship	X	X	.	X	X	
30	RELIG	Religion	.	.	X	X	.	
31	SCHOOL	School attendance	X	X	.	.	.	
32	EDATTAN	Educational attainment, international	X	X	.	.	.	
33	EDUCFR	Educational attainment, France	X	X	.	.	.	
34	EDUCNL	Educational attainment, Netherlands	.	.	X	X	X	
35	EEDATTA	Educational attainment, Europe	X	X	.	.	.	
36	EMPSTAT	Employment status	X	X	.	.	X	
37	EEMPSTA	Employment status, Europe	X	X	.	.	X	
38	OCCISCO	Occupation, ISCO	X	X	.	.	X	
39	OCC	Occupation, unrecorded	X	X	X	X	X	
40	INDGEN	Industry, general recode	X	X	X	X	X	
41	IND	Industry, unrecorded	X	X	X	X	X	
42	CLASSWK	Class of worker	X	X	X	X	X	
43	ECLASWK	Class of worker, Europe	X	X	X	X	X	
44	EMPSECT	Sector of employment	X	X	.	.	.	
45	EMPLNO	Number of employees		X	.	.	.	
46	HRSFULL	Full-time or part-time work	X	X	.	.	.	
47	LOOKJOB	Period seeking work	X	X	.	.	.	
48	PWRKFR	Region of work, France	X	X	.	.	.	
49	TRNWRK	Means of transportation to work or school	X	X	.	.	.	
50	MGRATEP	Migration status, previous residence	X	.				
51	MGRATE1	Migration status, 1 year	.	.	.	.	X	
52	MGRATEC	Migration status, last census	.	X	.	.	.	
53	MGCTRY4	Country of residence last census	.	X	.	.	.	
54	MIGFR	Region of residence at last census, FR	.	X	.	.	.	
	Person variables in 2006 sample not in 1999		11					
		Sum of Person Variables	57	49	17	18	21	17
	IPUMS constructed variables omitted from this listing		33	33	8	8	8	10

*****