SEMINAR: LES LUNDIS DE L'INED
6 JUNE 2005, INED

# INTERNATIONAL COMPARISONS BASED ON CENSUS MICRODATA (IPUMS): METHODS AND APPLICATIONS

Albert Esteve (Centre d'Estudis Demogràfics)
Robert McCaa, Steven Ruggles, Matt Sobek (University of Minnesota Population Center)

**Abstract.-** Census microdata are an invaluable resource for social science research. The Integrated Public Use of Microdata Series project (www.ipums.org) disseminates integrated census microdata to bona fide users over the Internet. Currently, harmonized census microdata from eight countries is available (Brazil, Colombia, China, United States, France, Kenya, Mexico, and Vietnam). During the next five years, IPUMS will distribute new microdata samples from most of the Latin American and European countries.

Building on the experience of previous integration projects, most of the technical challenges related to the international census microdata harmonization have been solved. With regard to variable harmonization, IPUMS utilizes composite coding schemes to guarantee maximum comparability across space and time without losing detail on the original variables. However, the theoretical implications of this initiative have no been yet fully explored.

Results from two research projects based on IPUMS data are presented to show the potential applications of the data base. The first one compares the effects of shrinking cohorts born during the 30s in France, United States and Spain on the marriage market and on nuptiality patterns. The second one compares levels of educational assortative mating in Mexico and Brazil.

**Résumé**.- Les micro-données des recensements sont une ressource précieuse pour la recherche en sciences sociales. Le projet *Integrated Public Use of Microdata Series* (www.ipums.org), consacré à la distribution par Internet des micro-données, offre à présent un accès aux micro-données harmonisées et intégrées de huit pays (Brésil, Colombie, Chine, Etats-Unis, France, Kenya, Mexique et Vietnam). Dans les 5 prochaines années, IPUMS sera prêt à distribuer les données de la plupart des pays d'Amérique Latine et d'Europe.

La longue trajectoire de ce projet a permis de surmonter la grande majorité des défis techniques qui se sont présentés. Pour l'harmonisation des variables, IPUMS a développé un système de codification multiple qui garantit en même temps une comparabilité maximale dans l'espace et dans le temps sans perdre les détails des variables d'origine. Cependant, la portée théorique de l'initiative n'a pas été encore suffisamment explorée.

Les résultats de deux recherches spécifiques sont présentés pour démontrer l'applicabilité et le potentiel de la base de données. Dans le premier cas, on compare l'effet de la dénatalité des années 30 en France, aux Etats Unis et en Espagne sur le marché matrimonial et sur les niveaux de nuptialité des générations nées dans ces années. Dans le deuxième, on compare le niveau d'homogamie éducative au Mexique et au Brésil.

**Introduction.-** 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. Census *microdata* provide information about individual persons and often families, households, and dwellings, usually in the form of one or more records per case, each consisting of a series of variables. For person records, age, sex, marital status, family relationship, place of birth, educational attainment, employment status, etc. are typical census microdata variables. Microdata are exceedingly useful because they allow researchers to interrelate any desired set of population and housing characteristics (Dale, Fieldhouse and Holsworth, 2000). The flexibility offered by microdata is essential for comparative research because aggregate tabulations produced by national statistical offices are usually not comparable across time or between countries. In the few countries where census microdata covering multiple census years have been easily available to researchers, these data are the most widely-used source for the study of large-scale economic and demographic transformations. Nevertheless, international comparisons based on census microdata are rarely attempted, partly because, for much of the world, census microdata are either unavailable or restricted, and are therefore seldom used (McCaa and Ruggles 2002).

This paper describes the IPUMS-International project, a consortium to anonymize, harmonize and distribute census microdata of a large number of countries. The paper is organized in three parts: an introduction to the IPUMS projects, a description of the IPUMS principles, and a practical application of its data to two research topics.

## PART I: THE IPUMS PROJECTS

**IPUMS-USA.-** In the United States and Canada census microdata have been available to researchers for almost forty years and have become an indispensable component of social science infrastructure. The Integrated Public Use Microdata Series (IPUMS-USA) is partly responsible for the widespread use of census microdata by social scientists studying the United States. IPUMS-USA, developed by Steven Ruggles, Matthew Sobek, and others at the Minnesota Population Center, makes census microdata freely available to scholars in harmonized format with comprehensive documentation through a user-friendly data access system (Ruggles and Sobek 1997; http://ipums.org/usa). Since its preliminary release in 1995, the IPUMS has become one of the most widely used demographic resources in the world. Over 6,000 researchers have registered to use the IPUMS data extraction system. The user base continues to expand rapidly, with approximately 2,500 new registered users during the past year alone. There have been prepared approximately 60,000 custom extracts of IPUMS data since May 1996 and are now processing approximately 2,800 data extract requests per month. This massive data distribution is beginning to bear fruit. Although the IPUMS has been available for only nine years, the bibliography lists more than twenty-six books, seventy-one dissertations, 207 published research articles, and hundreds of working papers, conference presentations, and research reports.

**IPUMS-International.-** In 1998 the IPUMS paradigm was extended to the censuses of Colombia. This pilot project, a collaboration with the Colombian National Statistical Office (DANE), was designed to demonstrate the feasibility of creating public use microdata for Latin America. Shortly after proposing the Colombia project, the National Science Foundation of the USA announced a special program for "Enhancing Infrastructure for the Social and Behavioral Sciences" that offered one-time funding for major new data improvement initiatives. A large-scale international project was proposed with two major components. The first step was to identify and preserve surviving machine-readable census microdata from around the world for the period 1960 to 2000. The second step was to select seven countries with broad geographical distribution and to clean, harmonize, document, and disseminate microdata for those countries using the same principles and methods that underlie the original IPUMS-USA database.

These two international projects, collectively known as IPUMS-International, have been an unqualified success. Both projects are now in their fifth year. In mid-2003, a Latin American imitative, including 16 Latin American countries with populations totaling one-half billion, was begun with funding by the National Institutes of Health.

The IPUMS team have created comprehensive inventory of known microdata, much of which is described in the *Handbook of International Historical Microdata* (Hall, McCaa, and Thorvaldsen 2000), and preserved microdata from over one hundred censuses. In the year 2002, the first preliminary group of harmonized census microdata samples for China (1982) Colombia (1964-1993), France (1962-1990), Kenya (1989-1999), Mexico (1960-2000), the United States (1960-1990), and Vietnam (1989-1999) was released. In June 2004, Brazil (1960, 1970, 1980, 1990, 2000) was also released. Over 50 million person records consisting of more than 40 variables are now available from the international web-site (http://www.ipums.org/international). Some fifty countries have now formally joined the IPUMS-International project (see Appendix 1).

During the first 33 months of operation, 766 applications were received, of which 39% were denied. The principal reason for rejecting an application is that the proposed research (as described by the applicant in the registration request—see Appendix A) does not seem to require access to the available microdata. The following statistics are derived from applications for access of the first 469 approved users of the IPUMS-International database.

Up to the current date, the following tables shows the distribution of the IPUMS users:

**Table 1. Country of residence and Countries of research interest (since August 2002)**

| **Country of residence** | **%** | | **Country/ies of interest** | **%** |
|---|---|---|---|---|
| USA | 72 | | Brazil (since Sept. 2004) | 4 |
| Canada | 4 | | China (since May 2003) | 11 |
| Switzerland | 3 | | Colombia | 13 |
| Brazil, Colombia, Kenya (total) | 8 | | France | 12 |
| France, Italy, Mexico, Spain, UK, Vietnam (total) | 6 | | Kenya | 12 |
| China (includes Hong Kong, etc.) | 1 | | Mexico | 20 |
| Australia, Germany | 1 | | USA (excludes IPUMS-USA) | 17 |
| 19 other countries (total) | 5 | | Vietnam | 11 |

**Table 2. User Profile: Institutional affiliation and Position**

| **Institutional affiliation** | **%** | | **Position** | **%** |
|---|---|---|---|---|
| University | 88 | | Student | 48 |
| Regional/International organization | 8 | | Researcher | 26 |
| National policy institute | 2 | | Professor | 21 |
| National statistical agency | 2 | | Other | 6 |

**Table 3. Academic discipline and Expected outcome**

| Academic discipline | ok% | | Expected outcome | % |
|---|---|---|---|---|
| Economics | 37 | | Teaching, B.A./M.A. thesis | 16 |
| Demography | 26 | | Paper, article, policy report | 10 |
| Sociology | 13 | | PhD dissertation | 9 |
| Public policy | 6 | | Book | 2 |
| History | 5 | | Enhance DHS/other survey | 6 |
| Other | 13 | | Other, Not mentioned | 57 |

**IPUMS-Europe**.- The National Institutes of Child Health and Human Development 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 fifty European censuses and micro-censuses from the 1960s to the present. The project will join the census microdata of Austria, Belarus, Bulgaria, the Czech Republic, France, Germany, Greece, Hungary, the Netherlands, Poland, Portugal, Romania, Slovenia, Spain, and the United Kingdom to the IPUMS-International integrated data series. Samples from Italy, Russia, and Turkey will also be incorporated pending written authorization from the national statistical agencies. These 18 countries represent over three-quarters of the population of Europe, and the resulting data series will offer excellent opportunities for research on a variety of topics, including analysis of population aging, economic transformation, demographic change, and international migration. This project goes substantially beyond the effort lead by the Population Activities Unit (PAU, Geneva) in three ways. First the PAU samples were limited to the 1990 round of censuses; so far only samples for 8 countries have been released. Second, the PAU effort was limited to standardizing codes for simple variables (age, sex, marital condition, employment status) and excluded more complex variables such as religion, family relationship, occupation, ethnicity, or language. Third, the samples were distributed on CDs, one per country, not as custom-tailored extracts based on the research needs of each individual researcher (Botev 2000).

Preliminary releases of harmonized European microdata may begin as early as 2006; the final integrated microdata series is scheduled for release in 2009. It will include between 50 and 60 datasets representing up to 70 million persons. A large number of housing, population, and economic variables are available for virtually every country, making the series particularly useful for a wide range of research projects.

The Centre d'Estudis Demogràfics (CED) at the Universitat Autónoma de Barcelona has received additional support. These funds provide for an inaugural workshop, to be held in 2005, at which census experts will discuss harmonization strategies to integrate European census microdata across space and time. The Sixth Framework Program will also support a three-year project to build a European web-based data dissemination extract site, which will make the European microdata and metadata more widely available for scholarly and educational research within the European region.

Despite the important contribution of the NIH funded project and the future contribution of the EU, further work remains to be done to fully capitalize on the potential of European census microdata. The CED has built a consortium of European Partners to accomplish the following important tasks:

- Extend the number of participating countries and obtain the latest censuses for those already participating.

- Develop classification schemes based on European standard nomenclatures.
- Build a network of participating countries to guarantee direct involvement during the harmonization process. This would be achieved through the celebration of workshops and seminars.
- Expand the array of complementary services to be provided along with the microdata (microdata handbook).
- Promote broader usage of the data by (i) organizing workshops for potential users from research institutes and universities and (ii) attending important demographic and social conferences.
- Promote substantive research at a cross-national level to explicitly demonstrate potential applications with the database.

## PART II: IPUMS PRINCIPLES

**Confidentiality protections.** The IPUMS-International differs from IPUMS-USA in one important respect: statistical confidentiality protections. The international 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 National Statistical Institute
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.

First, with regard to legal mechanisms, IPUMS-International partnerships are initiated only in countries where a memorandum of understanding signed by the official statistical agency authorizes a project. No work is begun—indeed no funds are solicited—for a project without prior signed authorization from each NSI.

Second, due to confidentiality restrictions, researchers must apply to become registered to use the system. Once registered, each user designs extracts that contain only the microdata for the countries, census years, sub-populations, sample densities and variables of interest to the specific research questions. It is noteworthy that approximately one-half of applications are denied access because of a failure to adequately satisfy one or another of the specified conditions.

Third, are the technical measures taken to ensure statistical confidentiality. In cases where the NSI requests that the MPC apply anonymization procedures, we implement the following technical protections (based on Thorogood 1999):

1. adopt sample size according to national norms or conventions;
2. limit geographical detail to administrative units with 20,000+ (or other minimum) population;
3. top and bottom code unique categories;
4. round, group, or band age as necessary;
5. suppress date of birth (report age);
6. suppress place of birth (<20,000 population);
7. suppress place of residence, work, study, and migration (<20,000 population);
8. systematically “swap” (recode) place of enumeration for a fraction of households;
9. randomly order households within administrative units;

10. and, conduct a sensitivity analysis once these are imposed to determine what additional measures may be required.

**Data Quality.** In addition to providing harmonized codes for variables and accompanying documentation, the IPUMS-International project is carrying out a variety of additional tasks to improve data quality. 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.** Harmonizing census data is not a new idea. First proposed in 1872 at the International Statistics Congress held in St. Petersburg, not much progress was made until the last half of the twentieth century. One of the signal achievements of the United Nations Statistics Division has been in the international harmonization of census concepts from the enumeration form to the publication of final tables. While incomplete, the effort has enjoyed widespread support by statistical agencies around the globe. Beginning in 1991, the IPUMS-USA project has worked to harmonize census data for the United States for the period since 1850, and IPUMS-International has capitalized on this experience.

The IPUMS-International projects adopt uniform coding schemes, nomenclatures and classifications, based where possible on the *Principles and Recommendations for Population and Housing Censuses* and other international standards such as:

- United Nations Statistics Division (1998) *Principles and Recommendations for Population and Housing Censuses*.
- United Nations Economic Commission for Europe (1999). *Recommendations for the 2000 Censuses of Population and Housing in the ECE Region* (Statistical Standards and Studies No. 49)
- UNESCO (1997) *The International Standard Classification of Education (ISCED 1997).*
- International Labor Office (1990) *International Standard Classification of Occupations (ISCO-88).*
- United Nations Statistics Division (1990) *International Standard Industrial Classification of All Economic Activities* (ISIC-88).

International 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 (Esteve and Sobek 2003).

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. Table 4 illustrates the harmonization of codes for the variable "employment status".

For example, the first digit of the variable for employment status is comparable across all samples. The second digit distinguishes (where appropriate) among the categories 'at work', 'not at work but had a work', and 'armed forces' for those individuals employed. The final digits provide additional detail with the unemployed category (such as number of month seeking for work).

The original codes in the census microdata are translated into a composite harmonized four-digit coding scheme. The range of concepts and coding schemes in this table hints at the complexities involved in developing a comprehensive system for a single variable. As more experience is gained by incorporating more countries and censuses, the table will surely be modified, but the basic structure of the composite coding scheme will remain. Thanks to the advice of experienced national consultants it is possible to readily identify problematic concepts and revise the harmonized codes accordingly. It is important to understand that no decisions are taken at the central integration center without comprehensive input by national experts who work as consultants to the project. This decentralized approach allows multiple projects to proceed simultaneously without fear of duplication or wasted effort.

The basic goal of our harmonization efforts is to simplify use of the data while losing no meaningful information. The IPUMS harmonization strategy has proven flexible enough to accommodate the integration of data across broad spans of time (the United States for 1850-1990) and space (Brazil, Colombia, France, Kenya, Mexico, the United States, and Vietnam).

**Data Dissemination.** The bulk of the web site documents is formed by the available samples and variables. Of particular note are the variable comparability discussions. These are designed to indicate where there are notable issues for interpreting a variable's codes for purposes of temporal and spatial comparison. In addition to these discussions, the web site contains the original census questionnaires and instructions so users can examine the full text from the original enumerations (www.ipums.org).

## Table 4. Harmonization Table for Employment Status

| Harmonized Codes and Labels | | Source Data Codes (selected samples) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| IPUMS-International | | Co | Co | Fr | Fr | Kn | Mx | Mx | US | Vn | Vn |
| Code | Label | 1964 | 1993 | 1962 | 1975 | 1999 | 1970 | 2000 | 1960 | 1989 | 1999 |
| 0000 | N/A | *,5 | B | * | B | BB | 0 | BB | 0 | B | B,1 |
| | ACTIVE (In Labor Force) | | | | | | | | | | |
| 1000 | EMPLOYED, not specified | 1 | | | | | | | | 1 | |
| 1100 | At work | | 4 | 1 | 1 | 1 | 1 | 10 | 10 | | |
| 1101 | At work, and 'student' | | | | | | | 14 | | | |
| 1102 | At work, and 'housework' | | | | | | | 15 | | | |
| 1103 | At work, and 'seeking work' | | | | | | | 13 | | | |
| 1104 | At work, and 'retired' | | | | | | | 16 | | | |
| 1105 | At work, and 'no work' | | | | | | | 18 | | | |
| 1106 | At work, public emergency | | | | | | | | 11 | | |
| 1107 | At work, family holding, not specified | | | | | | | | | | |
| 1108 | At work, family holding, not agricultural | | | | | 3 | | | | | |
| 1109 | At work, familiy holding, agricultural | | | | | 4 | | | | | |
| 1110 | Working and studying (France) | | | | | | | | | | |
| 1200 | Have job, not at work last week | | 3 | | | 2 | | 20 | 12 | | |
| 1300 | Armed forces | | | | | | | | 13 | | |
| 1301 | Armed forces, at work | | | | | | | | 14 | | |
| 1302 | Armed forces, not work last week | | | | | | | | 15 | | |
| 1303 | Military trainee (France) | | | 8 | 6 | | | | | | |
| 2000 | UNEMPLOYED, not specified | 2 | | | 3 | 5 | 2 | 30 | 20 | | |
| 2001 | Unemployed (Vietnam) | | | | | | | | | 4 | 5 |
| 2002 | Worked less than 6 months, permanent job | | | | | | | | | 2 | |
| 2003 | Worked less than 6 months, temporary job | | | | | | | | | 6 | |
| 2100 | Unemployed, experienced worker | | 1 | | | | | | 21 | | |
| 2101 | Seeking work, worked less than 3 months | | | 2 | | | | | | | |
| 2102 | Seeking work, worked 3 to 6 months | | | 3 | | | | | | | |
| 2103 | Seeking work, worked 6 to 12 months | | | 4 | | | | | | | |
| 2104 | Seeking work, worked more than 1 year | | | 5 | | | | | | | |
| 2105 | Seeking work, experience unspecified | | | 6 | | | | | | | |
| 2200 | Unemployed, new worker | | 2 | 7 | | | | | 22 | | |
| 3000 | INACTIVE (Not in Labor Force) | | | | | | | | 30 | | |
| 3100 | Housework | 3 | 6 | | | 10 | 3 | 50 | 31 | 6 | 2 |
| 3200 | Unable to work/disabled | 7 | 7 | | | 9 | | 70 | 32 | 7 | 4 |
| 3300 | In school | 4 | 5 | 9 | 5 | 7 | | 40 | 33 | 5 | 3 |
| 3400 | Retirees and living on rent | 8 | | | | | | 60 | | | |
| 3401 | Living on rent payments | | | | | | | | | | |
| 3402 | Retirees/pensioners | | 8 | | 4 | 8 | | | | | |
| 3500 | Elderly | 6 | | | | | | | | | |
| 3600 | No work available/discouraged | | | | | 6 | | | | | |
| 3700 | Inactive, other reasons | 9 | 0 | 0 | 0 | 11 | 4 | 80 | 34 | | 6 |
| 9000 | UNKNOWN/MISSING | | 9 | | | 0 | 9 | 99 | | | 9 |

**Note:** In the source data columns: a comma indicates more than one code was coded to the respective IPUMS-International value; an asterisk means programming logic was used; B indicates a blank in the source data.

## PART III: APPLICATIONS

The following applications have been chosen to demonstrate the potential uses of the IPUMS database to make international comparisons based on census microdata.

**Educational Homogamy in México and Brasil, 1970 – 2000[1].-** The first application explores and compares patterns of educational homogamy in Mexico and Brazil using harmonized samples of census microdata extracted from the IPUMS international database (Mexico 1970, 1990, 2000; Brazil 1970, 1980, 1991, 2000). The samples of individual records have been converted into samples of unions using SPLOC (spouse location), a household constructed variable provided by IPUMS to facilitate the linkage between partners. The final sample includes married and consensual couples formed by males and females aged 30 to 39 years old.

The creation of a comparable classification of education at international level poses the greatest challenge to this research. The countries examined do not share the same educational system. In terms of years of school, the Mexican system has its major divisions at 6 years (Primary), 9 (Lower Secondary), 12 (Secondary), and 16 (Tertiary). In the case of Brazil the divisions are 4 (Primary), 8 (Lower Secondary), 11 (Secondary), 15 (Tertiary). Sometimes censuses do not capture separately all these levels. This is the case of the Colombian 1993 census sample that combines some university education and a university degree in the same category because the census form was not designed to capture this distinction. For this reason, Colombia was not included in the analysis.

The IPUMS database offers two different educational variables: EDATTAIN (educational attainment) and YRSCHOOL (years of school).

EDATTAIN records the person's educational attainment in terms of the level of schooling completed (degree or other milestone). Thus a person attending the final year of secondary education receives the code for having completed primary schooling only. To create a consistent standard across countries, EDATTAIN applied the United Nations standard of six years of primary schooling, three years of lower secondary schooling, and three years of higher secondary schooling. For samples in which only individual years of study were originally reported, EDATTAIN grouped those years according to the 6-3-3 classification (see further discussion in http://www.ipums.org/international/descs/edattain_desc.shtml). However, this variable does not necessarily reflect any particular country's definition of the various levels of schooling in terms of terminology or the number of years of schooling. It is an attempt to merge -- into a single, roughly comparable variable -- samples that provide degrees, ones that provide actual years of schooling, and those that have some of both.

YRSCHOOL indicates the highest grade/level of formal schooling the person had completed, in years. YRSCHOOL accounts for the number of years of study, regardless of the track or kind of study (see more discussion at http://www.ipums.org/international/descs/yrschool_desc.shtml). YRSCHOOL has the advantage compared with EDATTAIN that allows accommodating and testing different classifications of education. In a dilemma about imposing a fixed classification based on the 6-3-3 schema mentioned above or creating country specific classifications that better reflect the school system of each country, the second option is preferred. For Mexico, individuals

[1] The results of this research will be presented at the XXV IUSSP conference, Tours 18-23 July. The complete paper will be available online by the time of the conference (http://iussp2005.princeton.edu/) The complete title of the paper is ‘Trends and patterns of educational homogamy in Mexico and Brazil: 1970-2000’ and the authors are Albert Esteve and Robert McCaa.

are classified into five groups: less than 6 years of school, 6-8, 9-11, 12-15, 16 or more. And there are also five groups for Brazil: less than 4 years, 4-7, 8-10, 11-14, 15 or more. In essences, these categories reflect the same educational boundaries in both countries: less the primary, primary, lower secondary, secondary and university completed.

Figures 1 and 2 show the distribution of males and females level of schooling by country and census year. A cursory examination of these figures reveals the dramatic educational expansion experienced by both countries in recent decades. The overall level of education is quite similar. Gender differences, though, are greater in Mexico than in Brazil, although they tend to reduce, as seen in year 2000.

Figure 3 reports the proportions of homogamous unions, those where husband and wife have the same level of schooling. The overall levels of homogamy decreases over time from 63,9% in Mexico 1970 to 45,7% in Mexico 2000 and from 74,8% in Brazil 1970 to 47,3% in Brazil 2000. For unions formed by partners of different educational attainment, we are interested in knowing whether heterogamic patterns are similar between men and women. For all years in Mexico and for 1970 and 1980 in Brazil, the proportion of female hypergamic unions is higher than the proportion of hypogamous unions. This pattern changes for Brazil 1990 and 2000.

Now, we ask whether the reduction of homogamous unions is affected by changes in the educational structure and whether the female hypergamic pattern is the product of unequal distribution of males and females schooling or rather results of marked social patterns. In order to control for changes in marginal distributions we relay on loglinear models, which distinguish patterns resulting from changes in marginal distribution, from patterns showing association between partners' schooling. Some of the results from the models are presented next. These results are net of the marginal distributions of husband's and wife's schooling[2].

The association between pairing with university completed is the highest of all and increases constantly over time in both countries (Figure 4). At a lower level, the same association pattern is observed between pairings with secondary education completed. Homogamy levels among individuals without primary completed remain constant over time. And homogamy levels among individuals with primary completed and lower secondary completed remain relatively low. Mexico and Brazil show the same patterns and trends but there is a difference in the degree of association, always higher in Brazil.

Finally, now, we analyze unions formed by partners of different educational attainment levels. The parameter (Figure 5) expresses that odds of women marrying up were higher than for men for all years except for Brazil 2000. The asymmetry parameters offer a synthetic but illustrative measure of the significance of the transformation of the female hypergamic pattern over time: in both countries the remarkable differences between the odds of men and women pairing with someone with more or less schooling shrink to the point of disappearing or even reversing, as seen in Brazil 2000. Again, the difference between Mexico and Brazil is that the odds of a woman marrying up are always lower in Brazil. To conclude, if each country had been studied independently, the same patterns would have been described: less overall homogamy, more homogamy among the higher educated, and less female hypergamic unions. Comparing both countries, however, some interesting remarks can be made: educational (social) differences are higher in Brazil while gender differences are higher in Mexico.

[2] A extended discussion of the models will be available in the final paper (see note 1).

**The Effects of Shrinking Cohorts on Family Formation: Historical Evidence in Spain, France, and United States in the 20th Century**[3]**.-** The effects of cohort sizes on family formation have been thoroughly studied, following Easterlin's seminal work, which identifies the labor market as the explanatory factor. In this application we test a different but converging hypothesis: with universal female marriage, women in shrinking birth-cohorts would marry younger and in greater proportions, that is, the marriage market would be the explanatory factor. As a result, age differences between spouses will increase. This kind of marriage squeeze should have rapid stimulating effects on female nuptiality, contrary to small effects where there is an excess of females. In two earlier works, the authors have developed the mechanisms of adjustment and tested them successfully for 20th Century Spain (Cabré 1993, 1994) , predicting from findings a reversal of fertility trends performed by the cohorts born after 1980. Using recent comparable census microdata, through IPUMS-International, the study is extended now to France and United States, where we seek to generalize the proof. These cases differ by their chronologies and by the imbalances of sexes at specific moments.

Little has been done to study the effects of marriage squeeze caused by sharp and lasting decreases in fertility, and occasional findings concerning these episodes have remained unexplained under this particular angle. However, this kind of marriage squeeze, characterized by the relative scarcity of women, would be of particular interest because of its rapid stimulating effects on nuptiality trends, contrary to its opposite, the female excess, which has been proven to have effects much lower than expected.

To identify shrinking cohorts, those affected by the marriage squeeze, we plot the number of births on a time scale. Between 1920 and 1945, at one point or another, all three countries experienced a significant decrease in the number of births. In the case of Spain (Figure 6), we focus on the effects produced by the decrease of births during the thirties. France has experienced two major contractions in the number of births in the course of the twentieth century (Figure 7). The first decrease, and most striking, occurred during the First World War. The second, less severe but longer in time, began in 1920 and last until 1940. This paper focuses exclusively on the effects of the second birth dearth. Finally, the US has also experienced two major contractions during the twentieth century (Figure 8). In this application we explore the decline occurring between 1924 and 1933, which was less sudden but longer than the French contraction.

The data used come from samples of census microdata of the 1991 Spanish Population Census (10%), US 1990 (1%), and France 1990 (5%) (the last two currently available at the IPUMS website: www.ipums.org). The availability of census microdata for the same period and for these countries offered an unique opportunity to carry this comparative research. Sex and age are the key variables for the analysis. These variables do not create problems of comparability. Unfortunately, none of the censuses report age of marriage. Thus, the effects of shrinking cohorts on the age of marriage can not be tested. Instead, age differences among spouses can be examined. Microdata supplied by IPUMs identifies spouses (SPLOC), which means that researchers may analyze husbands and wives according to their combined demographic characteristics.

[3] The results of this research will be presented at the XXV IUSSP conference, Tours 18-23 July. The complete paper will be available online by the time of the conference (http://iussp2005.princeton.edu/). The complete title of the paper is 'The effects of shrinking cohorts on family formation: historical evidence in Spain, France, and United States in the 20th Century' and the authors are Albert Esteve and Anna Cabré.

We use standard demographic methods to explore marriage patterns of targeted cohorts. We focus on the level (proportion of ever-married) and distribution (difference between male and female mean ages at marriage) of marriages.

In the case of Spain, proportions of never married went significantly down for women born between 1936 and 1945 and up for men of the same cohorts (Figure 9). Age differences between spouses increased for these women as well (Figure 10). In France, the male marriage squeeze had an effect on the proportion of males and females never married. The proportion of never married for men born between 1921 and 1939 went up and for women born during the same years went down (Figure 11). Age differences between spouses increased for French women born during the thirties (Figure 12). And, last, in the United States, similar patterns are found. Proportions of never married in the US are remarkably low compared to Spain and France. The proportion of never married for men born during the mid and late twenties and beginning of the thirties went up while female proportions remain low (Figure 13). The effects on the age gap are more visible. Age difference between spouses increased for women born between the mid and late twenties and beginning of the thirties (Figure 14).

To conclude, the expected effects of shrinking cohorts on family formation are verified in the three countries. However, when using a later census to explore the effects of an unbalanced marriage market in a distant past, some cautions have to be taken into account. Mortality, for instance, changes the distribution of the age differences between spouses as cohort of marriages age. Divorce, separation and remarriage also affect the observed distribution of marriages. And, finally, by comparing different cohorts at different ages we are in danger of mixing age and cohort effects. But in spite of these concerns, some rightful considerations can be made to illustrate, at least, the similarity between countries in their patterns.

**Conclusion.** Now that the construction of anonymized microdata data samples is becoming an increasingly widespread practice, harmonization of census microdata is an obvious next step to enhancing use. With the emergence of global standards of statistical confidentiality and the massive power of ordinary desktop computers, the only remaining obstacle is the integration of anonymized census microdata samples, which will, undoubtedly, encourage international comparisons, benefiting a wide range of research topics.

**References.**

Botev, Nikolai. 2000. PAU Census Microdata Samples Project. *In Handbook of International Historical Microdata for Population Research*, edited by Patricia Kelly Hall, Robert McCaa and Gunnar Thorvaldsen. Minneapolis: Minnesota Population Center, pp. 303-17.

Cabré, Anna, 1993. "Volverán tórtolos y cigüeñas" a Luís Garrido i Enrique Gil Calvo (eds.), *Estrategias familiares.* Madrid: Alianza Editorial, p. 113-131.

Cabré, Anna, 1994. "Tensiones inminentes en los mercados matrimoniales" a Jordi Nadal (ed.) *El mundo que viene.* Madrid: Alianza Editorial, p. 37-62.

Dale, A., Fieldhouse, E. and Holdsworth, C. (2000) *Analyzing census microdata*. Arnold: London.

Esteve, Albert and Matthew Sobek. (2003). Challenges and Methods of Census Harmonization. *Historical Methods* 36: 66-79.

Kelly Hall, P., McCaa, R. and Thorvaldsen, G., eds (2000) *Handbook of international historical microdata for population research.* Minnesota Population Center: Minneapolis. (Updated microdata inventory available at *www.IPUMS.org/ international/iiinventory2.html*.)

McCaa, Robert, and Steven Ruggles. 2002. The Census in Global Perspective and the Coming Microdata Revolution. In Vol. 13, *Nordic Demography: Trends and Differentials, Scandinavian Population Studies*, edited by J. Carling. Oslo: Unipub/Nordic Demographic Society, pp. 7-30.

Ruggles, S. (2000) 'The public use microdata samples of the U.S. census: research applications and privacy issues.' A report of the Task Force on Census 2000, Minnesota Population Center and Inter-University Consortium for Political and Social Research Census 2000 Advisory Committee. (Available at: *www.IPUMS.org/~census2000*.)

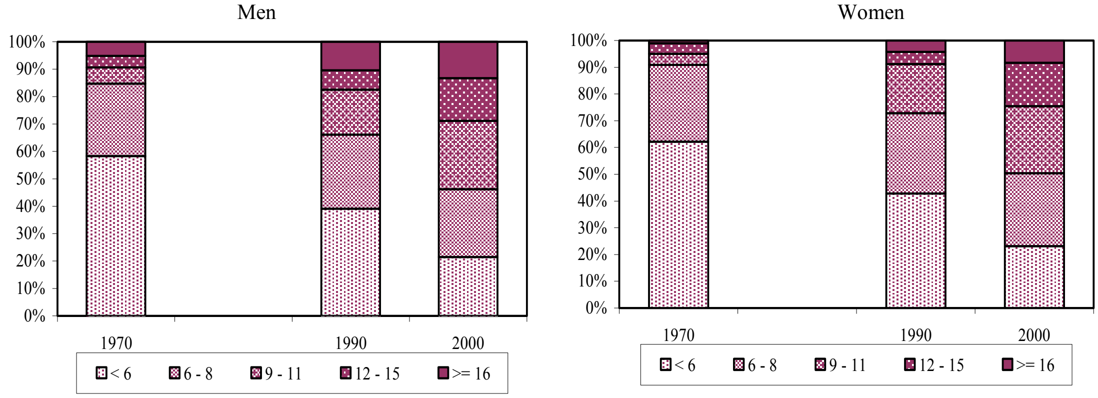
Ruggles, Steven, and Matthew Sobek, et. al. 1997. *Integrated Public Use Microdata Series: Version 2.0*. Minneapolis: Historical Census Projects, University of Minnesota.

Thorogood, D. (1999). 'Statistical Confidentiality at the European Level.' Paper presented at: Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Thessaloniki, March.

United Nations Statistics Division. (1998). *Principles and recommendations for population and housing censuses*. Department of Economic and Social Affairs, New York.

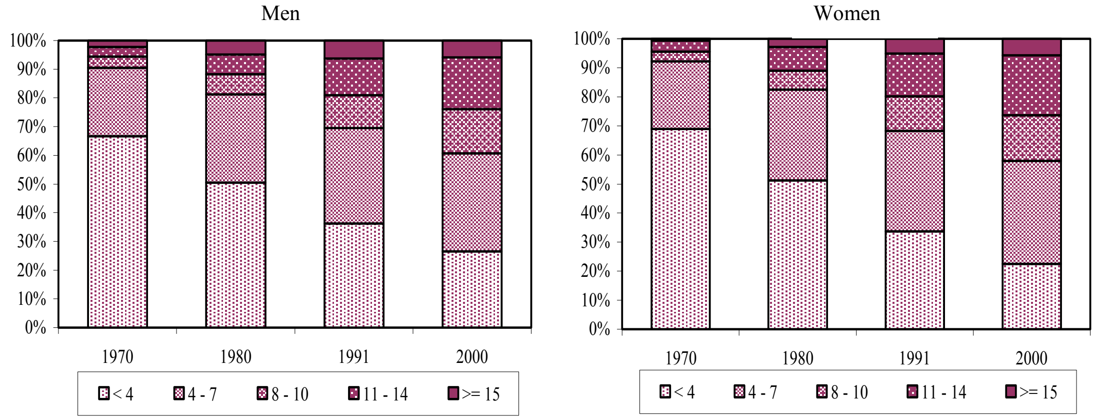
United Nations Economic Commission for Europe and Statistical Office of the European Communities. (1998). *Recommendations for the 2000 Censuses of Population and Housing in the ECE Region.* Statistical Standards and Studies, No. 49. New York and Geneva.

**Figure 1. Distribution of spouses by level of schooling, sex, and census year, Mexico**



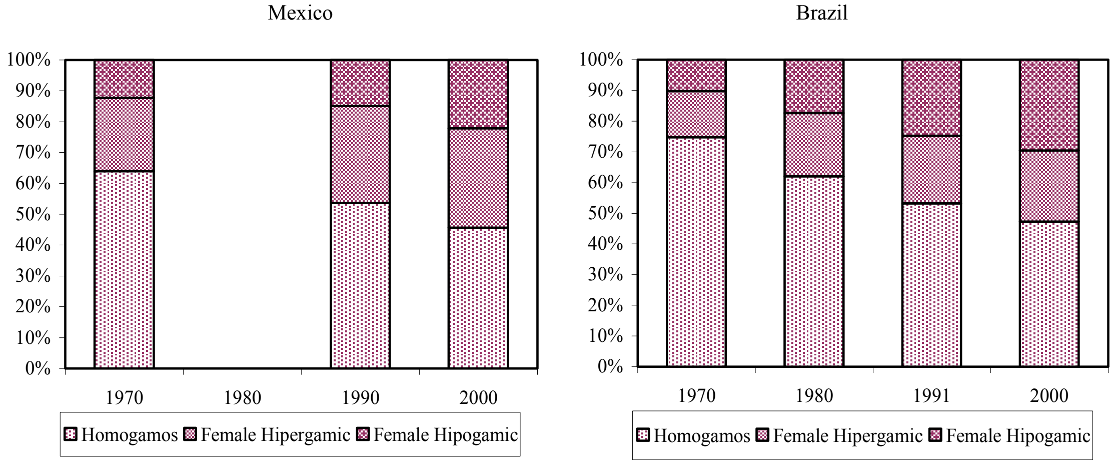
Source: Matthew Sobek, Steven Ruggles, Robert McCaa, et al., *Integrated Public Use Microdata Series-International: Preliminary Version 1.0*. Minneapolis: Minnesota Population Center, University of Minnesota, 2002.Estadística

**Figure 2. Distribution of spouses by level of schooling, sex, and census year, Brazil**



Source: Matthew Sobek, Steven Ruggles, Robert McCaa, et al., *Integrated Public Use Microdata Series-International: Preliminary Version 1.0*. Minneapolis: Minnesota Population Center, University of Minnesota, 2002.Estadística

**Figure 3. Distributions of unions by type, country and census year**



Source: Matthew Sobek, Steven Ruggles, Robert McCaa, et al., *Integrated Public Use Microdata Series-International: Preliminary Version 1.0*. Minneapolis: Minnesota Population Center, University of Minnesota, 2002.Estadística

**Figure 4. Log odds for homogamous pairings by level of schooling, country and census year**

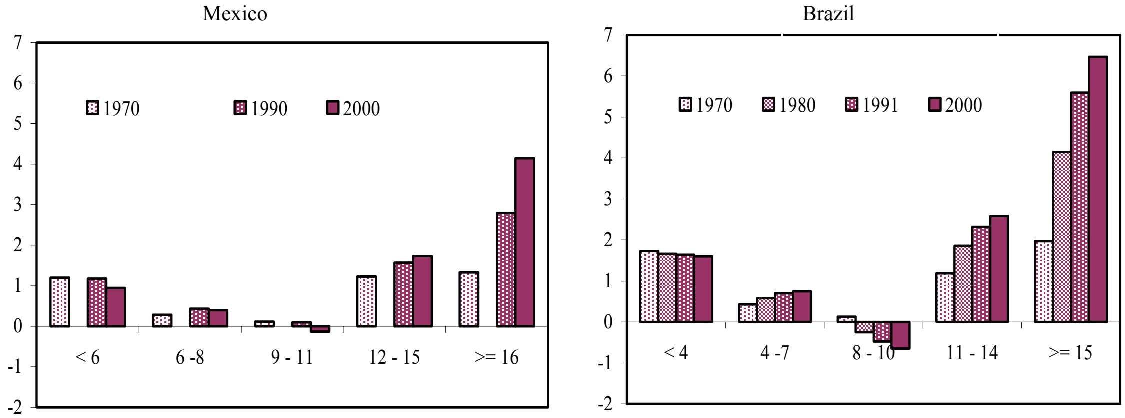


**Figure 5. Log odds for female hypergamic pairings by country and census year**

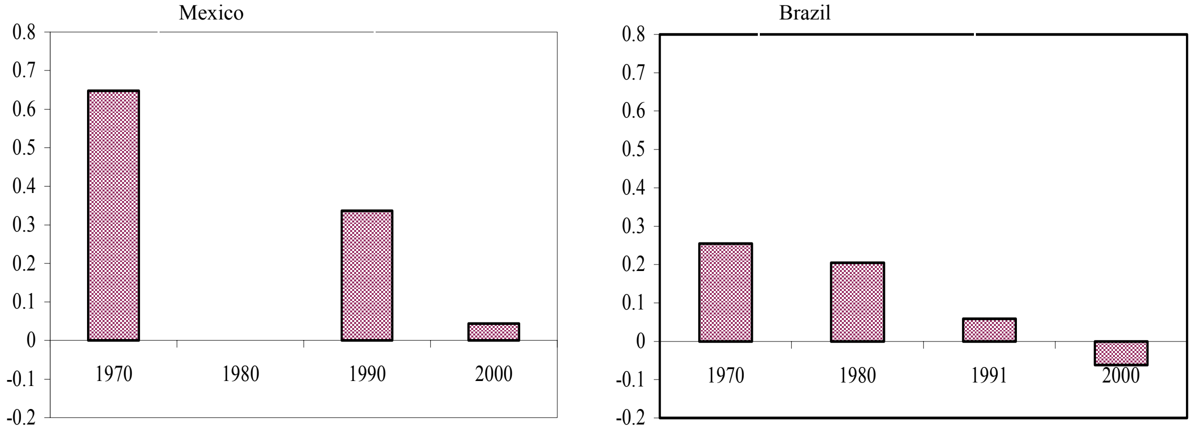
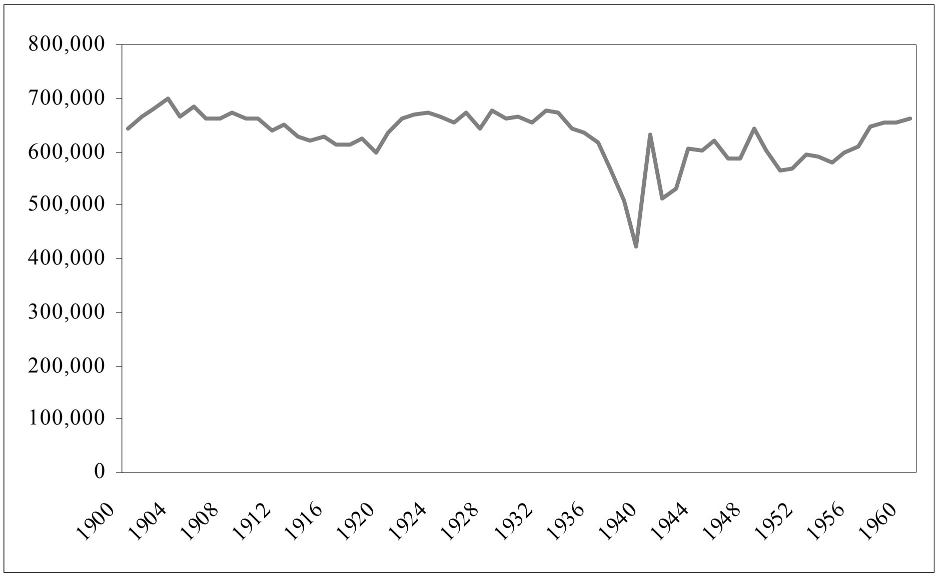
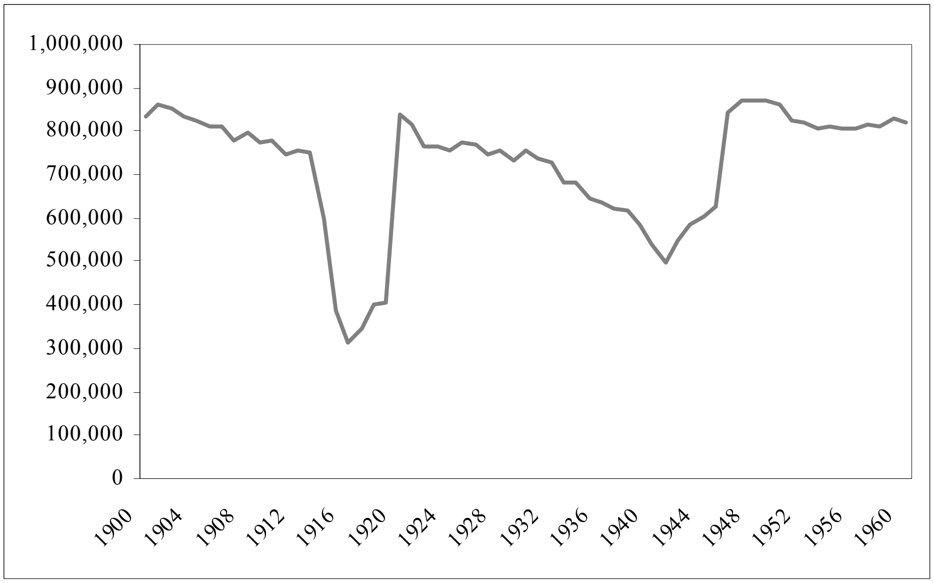


**Figure 6. Live Births Spain 1900 - 1960**



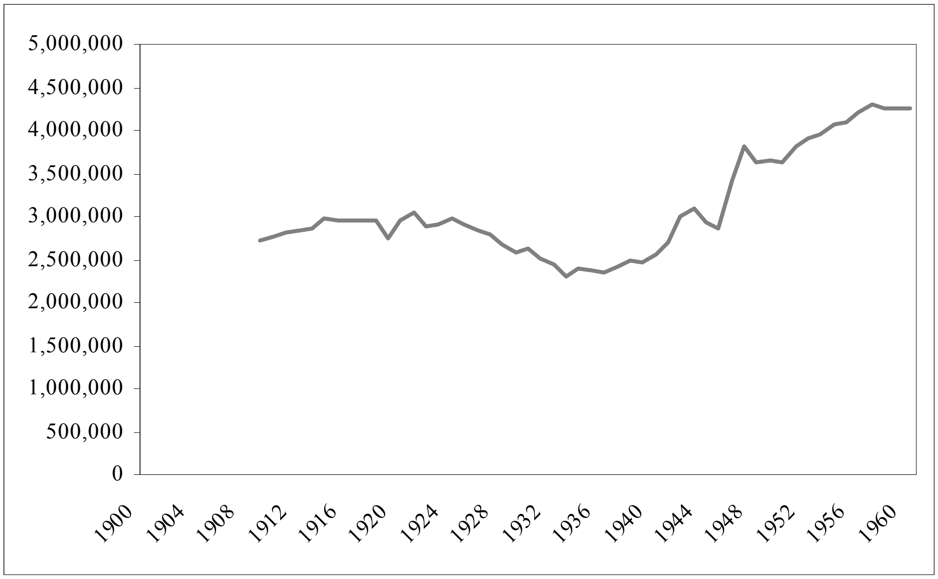
Source: Movimiento Natural de la Población Española (INE) (Vital Statistics)

**Figure 7. Live Births France 1900 – 1960**



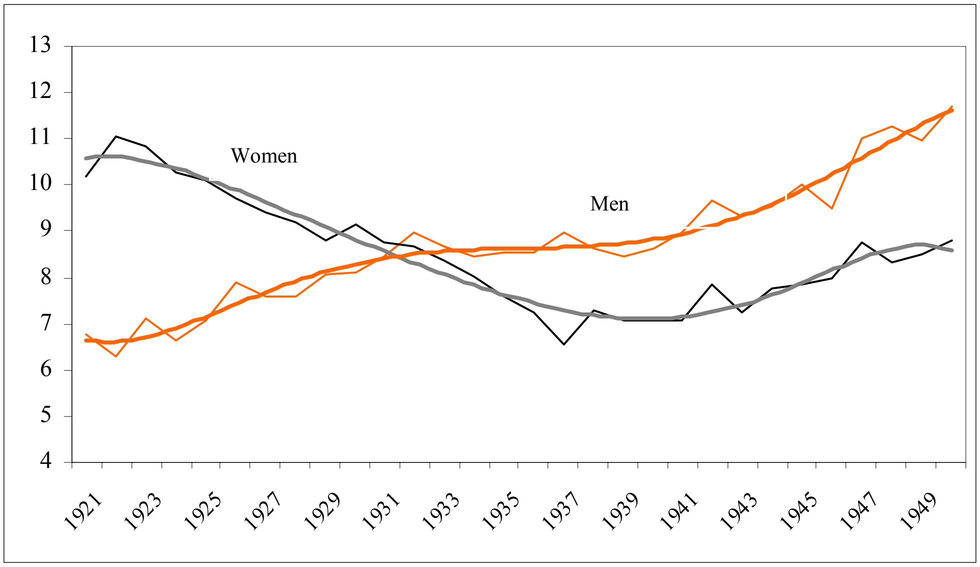
Source: INSEE, par Vallin, J. et Mesle, F. "Tables de mortalité françaises pour les XIXè et XXè siècle et projections pour le XXIè siècle", INED

**Figure 8. Live Births United States 1900 – 1960**



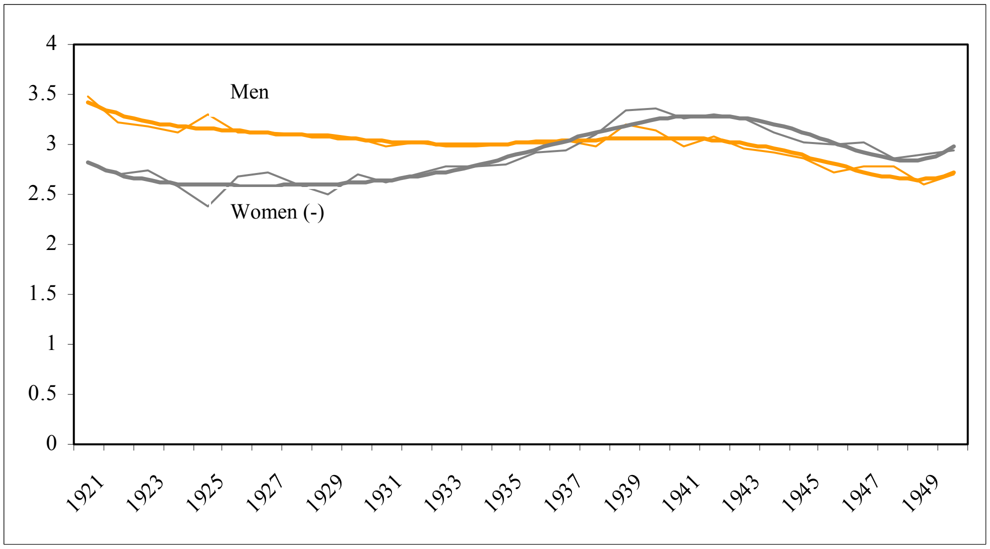
Source: U.S. Public Health Service, National Center for Health Statistics, Vital Statistics of the United States, 1993, vol. I, "Natality", Table 1□2; National Vital Statistics Reports, Vol. 48, No. 3, "Births: Final Data for 1998," Table 1.

**Figure 9. Proportion of never married by male and female cohort, Spain (Census 1991)**



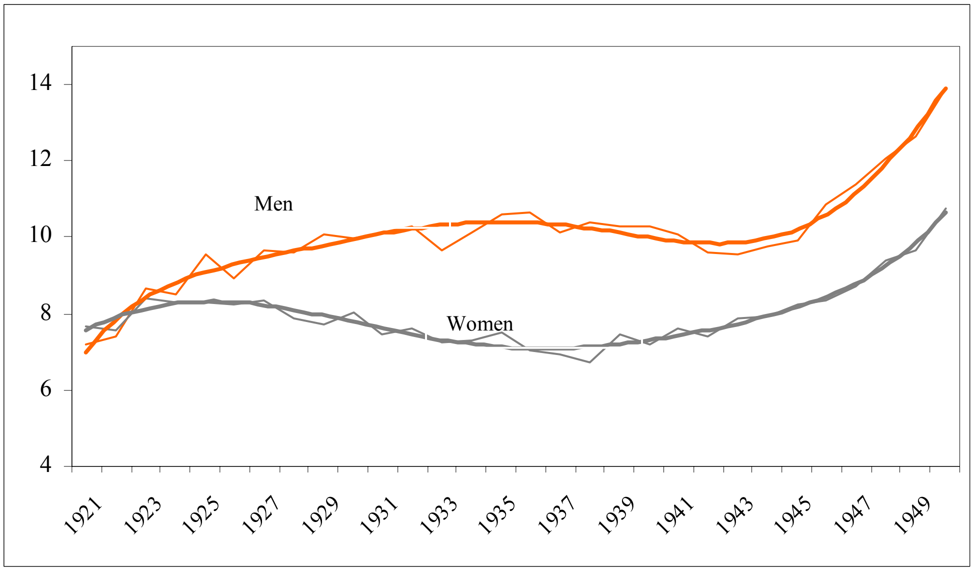
Source: Censo de Población de 1991, España, Instituto Nacional de Estadística

**Figure 10. Mean age difference between spouses by male and female cohort, Spain (Census 1991)**



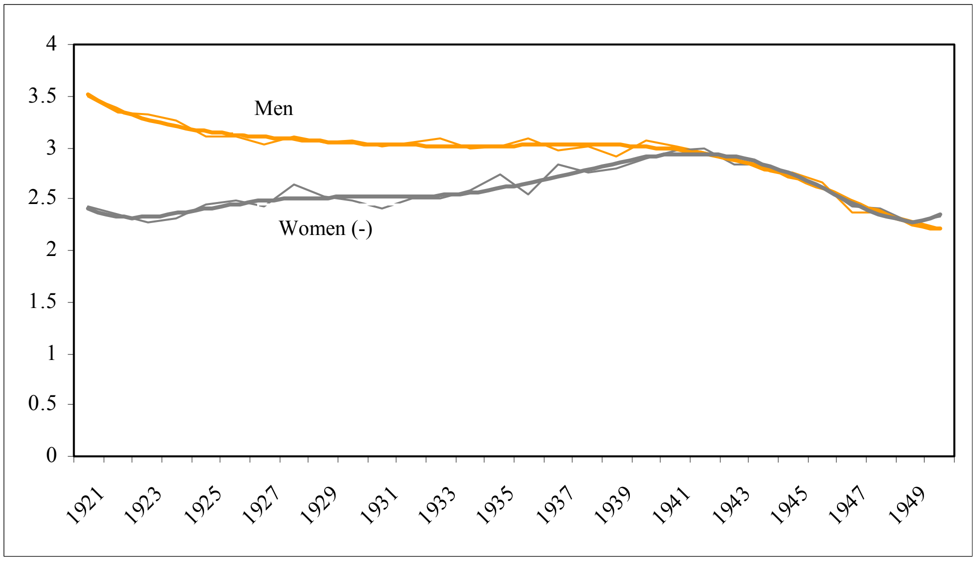
Source: Censo de Población de 1991, España, Instituto Nacional de Estadística

**Figure 11. Proportion of never married by male and female cohort, France (Census 1990)**



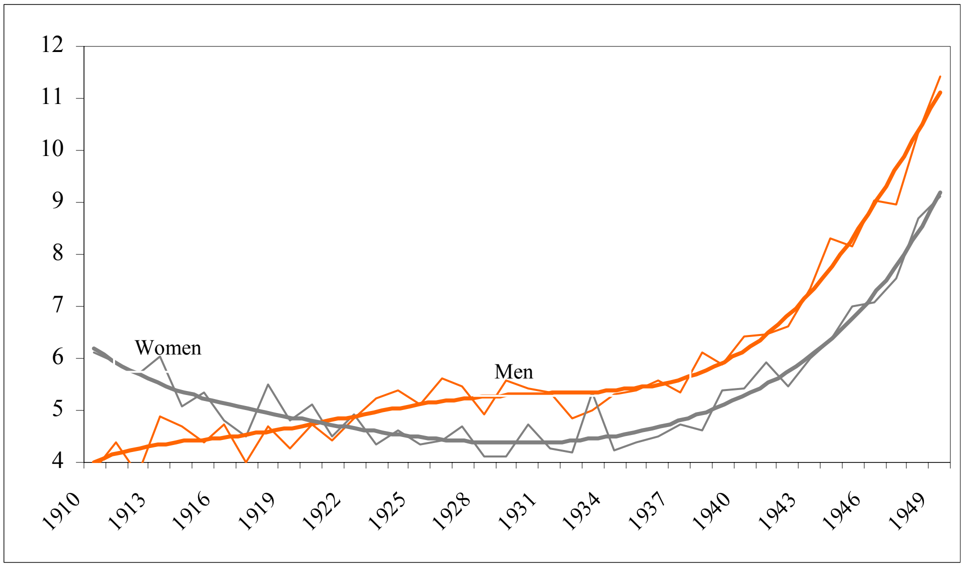
Source: Matthew Sobek, Steven Ruggles, Robert McCaa, et al., *Integrated Public Use Microdata Series-International: Preliminary Version 1.0*. Minneapolis: Minnesota Population Center, University of Minnesota, 2002.

**Figure 12. Mean age difference between spouses by male and female cohort, France (Census 1990)**



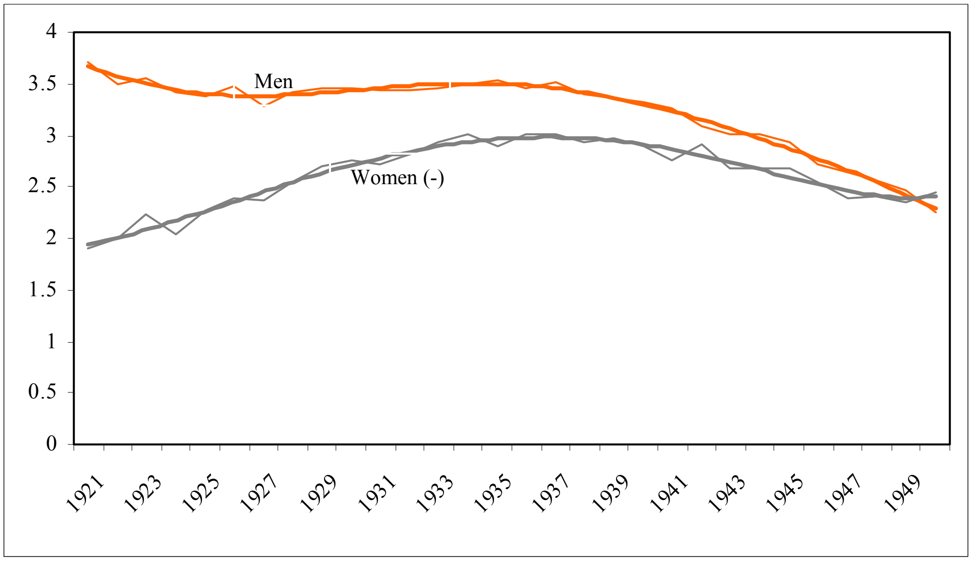
Source: Matthew Sobek, Steven Ruggles, Robert McCaa, et al., *Integrated Public Use Microdata Series-International: Preliminary Version 1.0*. Minneapolis: Minnesota Population Center, University of Minnesota, 2002.

**Figure 13. Proportion of never married by male and female cohort, United States (Census 1990)**



Source: Matthew Sobek, Steven Ruggles, Robert McCaa, et al., *Integrated Public Use Microdata Series-International: Preliminary Version 1.0*. Minneapolis: Minnesota Population Center, University of Minnesota, 2002.

**Figure 14. Mean age difference between spouses by male and female cohort, United States (Census 1990)**



Source: Matthew Sobek, Steven Ruggles, Robert McCaa, et al., *Integrated Public Use Microdata Series-International: Preliminary Version 1.0*. Minneapolis: Minnesota Population Center, University of Minnesota, 2002.Estadística

**Appendix 1. IPUMS-International Country Partners and Census Microdata February 1, 2005**

Key: * = microdata for all census years entrusted to project
Year = census conducted;
**Bold year** = microdata survive; m = microcensus

| | **Place** | **2000s** | **1990s** | **1980s** | **1970s** | **1960s** |
|---|---|---|---|---|---|---|
| **Phase I, 1999-2004** | | | | | | |
| * | Brazil | **2001** | **1991** | **1980** | **1970** | **1960** |
| | China (only '82 so far) | **2000** | **1990** | **1982** | | 1964 |
| * | Colombia | | **1993** | **1985** | **1973** | **1964** |
| * | France | **1999** | **1990** | **1982** | **1975** | **1968, 62** |
| * | Kenya | **1999** | **1989** | **1979** | **1969** | |
| * | Mexico | **2000** | **1990** | 1980 | **1970** | **1960** |
| * | United States | **2000** | **1990** | **1980** | **1970** | **1960** |
| * | Vietnam | | **1999** | **1989** | 1979 | |
| **Phase II, 2004-2008** | | | | | | |
| **Asia and the Pacific** | | | | | | |
| | Armenia | **2001** | | 1989 | 1979 | |
| | Bangladesh | **2001** | **1991** | **1981** | **1974** | 1961 |
| * | Cambodia | | **1998** | | | |
| | Georgia | **2002** | | 1989 | 1979 | |
| | Iran | | **1996** | **1986** | **1976** | **1966** |
| | Iraq | | **1997** | 1987 | 1977 | 1967 |
| | Israel | | **1995** | **1983** | **1972** | **1961**, 67 |
| * | Malaysia | **2001** | **1991** | **1980** | **1970** | 1960 |
| * | Mongolia | **2000** | | 1989 | 1979 | |
| | Pakistan | | **1998** | **1981** | **1973** | 1961 |
| * | Palestinian Authority | | **1997** | | | |
| * | Philippines | **2000** | **1990** | **1980** | **1970** | **1960** |
| | Tajikistan | **2000** | | 1989 | 1979 | |
| | Turkmenistan | | **1995** | 1989 | 1979 | |
| **Europe** | | | | | | |
| | Austria | **2001** | **1991** | **1981** | **1971** | 1961 |
| * | Belarus | | **1999** | 1989 | | |
| | Bulgaria | **2001** | **1992** | 1985 | 1975 | 1965 |
| | Czech Republic | **2001** | **1991** | 1980 | 1970 | 1961 |
| | Germany | **2001m** | **1991m** | **1987, 81** | **1970, 71** | 1961 |
| | Greece | **2001** | **1991** | **1981** | **1971** | 1961 |
| * | Hungary | **2001** | **1990** | **1980** | **1970** | |
| | Ireland | **2001** | **1991** | 1981 | 1971 | 1961 |
| | Netherlands | **2001m** | | | **1971** | **1960** |
| pending | Poland | **2001** | | **1988** | **1978**, 70 | 1960 |
| | Portugal | **2001** | **1991** | **1981** | 1970 | 1960 |
| * | Romania | **2001** | **1992** | | 1977 | 1965 |
| pending | Russia (-1989 USSR) | **2002** | **1994m** | **1989** | 1979 | 1970 |
| | Slovenia | **2001** | **1991** | 1981 | | |
| * | Spain | **2001** | **1991** | **1981** | 1970 | 1960 |
| pending | Turkey | **2000m** | **1990** | 1980, 85 | **1970, 75** | 1960, 65 |
| | United Kingdom | **2001** | **1991** | 1981 | 1971 | 1961 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| **North America** |  |  |  |  |  |  |
|  | Canada | **2001** | **1991,** 96 | **1981,** 86 | **1971,** 76 | 1961, 66 |
| * | Costa Rica | **2000** |  | **1984** | **1973** | **1963** |
| * | El Salvador |  | **1992** |  | **1971** | 1961 |
|  | Guatemala | 2003 | **1994** | **1981** | **1973** | **1964** |
|  | Honduras | **2000** |  | **1988** | **1974** | **1961** |
|  | Nicaragua |  | **1995** |  | **1971** | 1963 |
| * | Panama | **2000** | **1990** | **1980** | **1970** | **1960** |
| **South America and Caribbean** |  |  |  |  |  |  |
|  | Argentina | **2001** | **1991** | **1980** | **1970** | **1960** |
| * | Bolivia | **2001** | **1992** |  | **1976** |  |
| * | Chile | **2002** | **1992** | **1982** | **1970** | **1960** |
|  | Dominican Republic | 2003 | 1993 | **1981** | **1970** | 1960 |
| * | Ecuador | **2001** | **1990** | **1982** | **1974** | **1962** |
| * | Paraguay | **2002** | **1992** | **1982** | **1972** | **1962** |
| * | Peru |  | **1993** | **1981** | 1972 | 1961 |
| * | Puerto Rico | **2000** | **1990** | **1980** | **1970** | 1960 |
| * | Venezuela | **2001** | **1990** | **1981** | **1971** | **1961** |
| **Africa** |  |  |  |  |  |  |
|  | Egypt |  | **1996** | **1986**, 81 | 1976 | 1964 |
|  | Madagascar |  | **1993** |  | 1975 | 1966 |
|  | Malawi |  | **1998** | **1987** | **1977** | 1966 |
| * | South Africa | **2001** | **1996,** 91 | 1985, 80 | 1970 | 1960 |
|  | Uganda | **2000** | **1991** | **1980** |  | 1969 |
|  |  |  |  |  |  |  |
| **Datasets per Census Round (n)** |  | **43** | **52** | **37** | **36** | **18** |