WP.38
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (iii): Applications (including practical implementation of SDC methods, actual issues within NSIs and software)

INTEGRATED EUROPEAN CENSUS MICRODATA (IECM) SAMPLES: ENHANCING THE STUDY OF AGEING WITH HIGH PRECISION OVER-SAMPLES OF THE OLDEST-OLD

Supporting Paper

Prepared by Albert Esteve, Joan Garcia and Jeroen Spijker (Universitat Autònoma de Barcelona, Spain); and Robert McCaa (University of Minnesota, United States of America)

Integrated European Census Microdata (IECM) Samples: Enhancing the study of ageing with high precision over-samples of the oldest-old

Albert Esteve*, Joan Garcia*, Jeroen Spijker*, Robert McCaa**

* Centre d'Estudis Demogràfics, Campus Universitat Autònoma de Barcelona, Bellaterra, 08193, SPAIN aesteve@ced.uab.es, jgarcia@ced.uab.es, jspijker@ced.uab.es

** Minnesota Population Center, University of Minnesota

Abstract: A breakthrough in the tradeoff between privacy and data quality has been achieved for restricted access to anonymized population census microdata samples of Europe. As of September 2007, the IECM website, in partnership with the Minnesota Population Center, offers integrated microdata for 22 censuses, totaling more than 31,2 million person records, with 7 European countries represented. Over the next two years, the European collaboratory led by the Centre d'Estudis Demogràfics and the Minnesota Population Center, with major funding by the 6th Framework, United States National Science Foundation and the National Institutes of Health, will disseminate samples for more than 25 additional censuses. Thanks to high precision samples of one, five and even ten percent, much research on the ageing of populations can be accomplished. Nevertheless for studies of the oldest-old, over-samples of the elderly will often be required. The paper examines basic statistics on headship rates in the samples currently available and illustrates some of the limitations that can best overcome by oversamples for elderly populations.

1 The Integrated European Census Microdata (IECM) project

A vast quantity of census microdata covering Europe in the period since the 1960s survives in machine-readable form. Most of these data, however, remain inaccessible to researchers. The Centre d'Estudis Demogràfics located at the Universitat Autònoma de Barcelona along with other leading European research centers is involved in an international initiative, lead by the Minnesota Population Center (MPC), to create a harmonized and documented data series based on samples of over fifty Western and Eastern European censuses. The project is funded by the National Institutes of Health (NIH), a funding agency of the United States of America. In collaboration with the national statistical agencies of each country, we have negotiated redistribution agreements for the censuses of 17 European countries: Austria, Belarus, Bulgaria, Czech Republic, France, Germany, Greece, Hungary, Italy, the Netherlands, Portugal, Romania, Slovenia, Spain, Switzerland, Turkey and the United Kingdom (see Table 1). Combined, these countries account for over sixty percent of the population of Europe. The European census microdata series has an important chronological dimension. In all seventeen countries, the data span twenty years; for eight countries, thirty years; and for three countries, forty years. With over fifty million persons residing in over twenty million households, the integrated European dissemination system offer broader chronological scope and greater sample densities than any alternative data source. In most cases, the censuses are also the most representative source of information available about national populations.

Grants from the National Institutes of Health and the National Science Foundation of the United States cover the costs of finding and preserving microdata and documentation, negotiating dissemination agreements, developing data cleaning and sampling procedures, creating data conversion and dissemination software, and establishing design protocols for data and documentation. Additional funding from the Sixth Framework Programme cover the costs of

coordination, dissemination and harmonization tasks based in Europe. On March 2007, the first European samples were released. Data access is provided for Belarus (1999), France (1962, 1968, 1975, 1982, 1990), Greece (1971, 1981, 1991, 2001), Hungary (1970, 1980, 1990, 2001) Portugal (1981, 1991, 2001), Romania (1991, 2002) and Spain (1981, 1991, 2001). In 2008, samples for Austria, Netherlands and the UK will be launched.

Table 1. IECM-Europe: microdatasets entrusted by country, subsample precision and design For current data availability, see: http://www.iecm-project.org .

Datasets entrusted by subsample precision			Country	Sub sample design	2000s	1990s	1980s	1970s	1960s
10%	~5%	<=4%							
4			**Austria**	IPUMS	**2001**	**1991**	**1981**	**1971**	1961
1			**Belarus**	IPUMS		**1999**	1989	1979	1970
			Bulgaria (in process)		**2001**	**1992**	**1985**	1975	1965
	2		**Czech Republic**	IPUMS	**2001**	**1991**	**1980**	**1970**	1961
	5		**France** ('99 in process)	IPUMS	**1999**	**1990**	**1982**	**1975**	**1968, 2**
1			**Germany** (in process)	IPUMS	**2001m**	**1991m**	**1987, 1**	**1970, 1**	1961
4			**Greece**	IPUMS	**2001**	**1991**	**1981**	**1971**	1961
	4		**Hungary**	IPUMS	**2001**	**1990**	**1980**	**1970**	
			Italy (in process)		**2001**	**1991**	1981	1971	1961
		3	**Netherlands**		**2001m**			**1971**	**1960**
			Poland (negotiating)		**2001**		**1988**	**1978**, **0**	1960
	3		**Portugal**	IPUMS	**2001**	**1991**	**1981**	1970	1960
2			**Romania** ('77 recovered)	IPUMS	**2001**	**1992**		**1977**	1965
			Slovenia (in process)		**2001**	1991	1981		
	3		**Spain**	IPUMS	**2001**	**1991**	**1981**	1970	1960
			Switzerland (in process)	IPUMS	**2000**	**1990**	**1980**	**1970**	1960
			Turkey (in process)		**2000**	**1990**	1980, 5	1970, 5	1960, 5
		2	**United Kingdom**		**2001**	**1991**	**1981**	**1971**	1961

Note: **bold country** = Agreement signed between University of Minnesota and National Statistical Authority
Year = census; **Bold year** = microdata survive; * = 100% microdata entrusted to IECM; m = microcensus
IECM systematic subsample design for private households: every n^{th} household stratified by enumeration district.

The database allows social scientists to make comparisions across European nations and opens the doors to European Statistical offices for transnational access to one of the most valuable assets for social science research. The database is an exceptionally valuable resource for researchers working on a broad range of topics. The large samples offered by European census microdata, are an invaluable resource compared to other sources such as demographic and labor forces surveys, which often offer greater subject coverage and detail, but less sample density, chronological depth, or geographic coverage. Thanks to high precision samples of one, five and even ten percent (see Table 1), much research on the ageing of populations can be accomplished. Nevertheless for studies of the oldest-old, over-samples of the elderly will often be required.

A comprehensive array of protections is in place to guarantee the privacy and statistical confidentiality of census microdata samples incorporated into the IECM database (McCaa, Esteve, 2006). These protections involve three elements:

1. legal: dissemination agreements between the University of Minnesota and each participating Official Statistical Institute

2. administrative: licenses between the University of Minnesota and each user, specifying conditions and restrictions of use
3. technical: perturbations of the data (swapping, recoding) to make exceedingly unlikely the identification of individuals, families or other entities in the data. Technical measures have the additional benefit that any assertion of absolute certainty in identifying anyone in the data is false.

The IECM project, by disseminating only integrated, anonymized microdata and restricting access to licensed academic users, shifts the risk-utility curve sharply upward. This approach provides access to microdata of high utility at the same time that confidentiality risks are minimized.

2 Enhancing the study of ageing with high precision over-samples of the oldest-old

2.1. Aging in Europe: the continuing need for better research and data

Europe is the continent with the highest proportion of elderly population, both in terms of the old (60-79) and oldest-old (80+). While the latter group only encompased 1,1% of the population in Europe in 1950, today they account for 3,5% and it is estimated that in 2020 5,1% will be 80 years or older and in 2050 9,6% (United Nations 2007). However, it is not only that a segment of the population that is growing in both absolute and relative numbers because of falling birth rates, they are also doing better than ever in terms of lower levels of mortality (United Nations 2007) and morbidity (Murray and Lopez 1996). As a consequence of such large increases in people surviving to older ages, there are now much higher demands in health care (that are mainly for care, not cure), social welfare services, pension funds and housing specifically for elderly needs, etc (European Commission 2005) that will only increase in future.

Besides the aforementioned growing welfare and housing needs, increased longevity has also had its impact on the oldest-old household composition and population structure by sex and marital status (more are still married, but there is also a growing group of divorced elderly). Changes in both household composition and marital status are, however, not only related to demographic factors, as changes in norms and values have made it both legally possible and socially acceptable for the widowed and divorced to remarry or to cohabit without formal marriage, whereby even among the oldest-old repartnering is no longer considered a taboo or something they themselves wouldn't consider given their age, at least for those who are in relatively good health (e.g. Lopata, 1996).

Finally, due to its cultural, historical and political diversity, there are still many differences between and even within the European countries as to both the household composition and welfare and household demands of the oldest-old population. One example of intra-European differences is with respect to the proportion of the oldest-old population that is institutionalised or that live in semi-autonomous elderly residential homes, which is more common in northern countries, while in the southern countries it is the family that is expected to take on the majority

of caring responsibilities of ailing or disabled elderly family members. In a similar manner, many rural areas, particularly those far away from mayor urban areas, lack the infrastructure for the supply of sufficient resources for the oldest-old, especially in the area of health care. Not surprisingly, population ageing and the continual increase in life expectancy, including of the oldest-old, has also excited a wide interest among researchers to study theories of ageing (e.g. Olshansky et al. 1990, 2005; Yashin 2001), to produce population projections of the oldest-old (e.g. Boleslawsky and Tabeau 2001; United Nations 1997, 2007), as well as to study specific implications of an ageing society in terms of changing household structure, changes in relationship forming, social policy, health care, disability and housing (see European Commission 2005).

However, a common problem is that there are few sources that permit a better and more accurate description of both current, past and changing household characteristics of the elderly, their living arrangements and living conditions. Moreover, the sample size for studies of the oldest-old is often small which undermines not only the research possibilities but also the reliability and validity of the obtained results. This is why we advocate greater use of census microdata as a means to study population characteristics of the elderly like their household structure, living arrangements and socioeconomic situation as it usually contains the entire population or a large sub-sample.

At the same time, we also advocate for future censuses where only low sample densities are made available for researchers that oversampling is applied to the oldest-old as this will provide the scientific community with an improved statistical basis for their research when the conventional sample size is not enough. Oversampling has been used in the past as a way to obtain a larger representation of small population groups that were of special interest for a particular study but where there were relatively few cases and therefore would cause larger statistical uncertainty in the results. For example, oversampling was recommended by Coleman et al. (2000) in order to better explore the mechanisms underlying some of the trends, patterns, and relations found in quantitative work on different types of relationship formations, that although still uncommon among the elderly today, is increasing in importance. This especially includes non-traditional living arrangements like Living Away Together or non-marital cohabitation after widowhood and divorce. In order to better understand the meanings of such experiences of people in the different cultural contexts (as we stress in this paper the importance of doing international comparisons) we could gain considerable insight from such qualitative approaches as in-depth interviewing when particular sub-sets of the population (in this case the elderly) is oversampled.

2.2. The IECM census samples and the oldest-old: Why are larger sample sizes needed?

To illustrate some of the limitations of the IECM census sample densities with regard to the oldest-old population, we examine basic statistics on headship rates in the samples currently available, and which can be best overcome by oversampling. Figure 1 shows the width of the 95% confidence intervals, that is, the range in percentage points between the lowest and the highest values. The confidence intervals have been calculated in the usual manner.

$$Z = 1{,}96 * \sqrt{\frac{N-n}{N} * \frac{p*(1-p)}{n}}$$

Where *N* equals to the total population; *n* to the sample size, and *p* to the observed rate.

We considered the most recent sample available of actual IECM countries that provided regional level data[1] (for this reason Hungary is excluded). For each sample we estimated sex specific headship rates by age, age and marital status, and age, marital status and region. With regard to the first case, we simply calculated confidence intervals by single age except for those aged above 95 years, which we grouped into an open ended age group. Secondly, we computed the width of the confidence intervals for each combination of age and marital status (single, married, divorced and widowed). The results that are given in the figure are weighted averages of the four marital statuses by age. A similar approach was applied to the age, marital status and regions combinations[2]. The regional dimension was included to test the possibilities of carrying out sub-national analyses.

As to be expected, confidence intervals increase with age and are wider for men than for women. Due to the specific old age population sizes, confidence intervals really begin to increase between ages 75 and 85. At the national level, with or without the inclusion of marital status, the range of the confidence intervals rarely exceeds 10 percentage points, except at very old ages. The lowest variations of headship rates are found among French and Spanish women, while Portuguese men generally show the highest range in age-specific confidence intervals. The impact of including a regional dimension on the width of the confidence intervals depends on the number of regions available in the sample for each country. By definition, the confidence interval is larger when region is included. In general, analysis by region increases the width of the confidence intervals to above 10 percentage points and among the oldest-old to more than 20 percentage points.

As noted above, one way to overcome the statistical limitations of small sample sizes that particularly affect old age groups is to over-sample this subset of the population. In our example, we show the effects of increasing sample sizes two-, three-, four- and five-fold, on the confidence interval widths for the sex specific headship rates by age, marital status and region (see Figure 2). As is clearly shown in the figures, increasing sample size dramatically shrinks the width of the confidence intervals, especially for those countries with either a small elderly population and/or small sample densities. However, due to the high variability in confidence intervals by age, sex and country it would be better to "customize" the specifications of high precision samples for the oldest-old. For example, according to the results for France and Portugal, oversampling would need to be performed across all ages, but in the case of Belarus and Rumania, perhaps only for ages 80 and over. One should bear in mind, however that this applies to age, sex and marital status headship rates that we used as an illustrative example. Other types of variables might yield smaller or even greater confidence intervals.

[1] Belarus 1999 (10%), France 1990 (4,2 %), Greece 2001 (10%), Portugal 2001 (5%), Romania 2002 (10%), and Spain 2001 (5%).

[2] Belarus 1999: 6 regions; France 1990: 22 regions; Greece 2001: 54 regions; Portugal 2001: 7 regions; Romania 2002: 8 regions; Spain 2001: 52 regions.

Figure 1. Width of the confidence interval (in % points) of headship rates by age, marital status and region

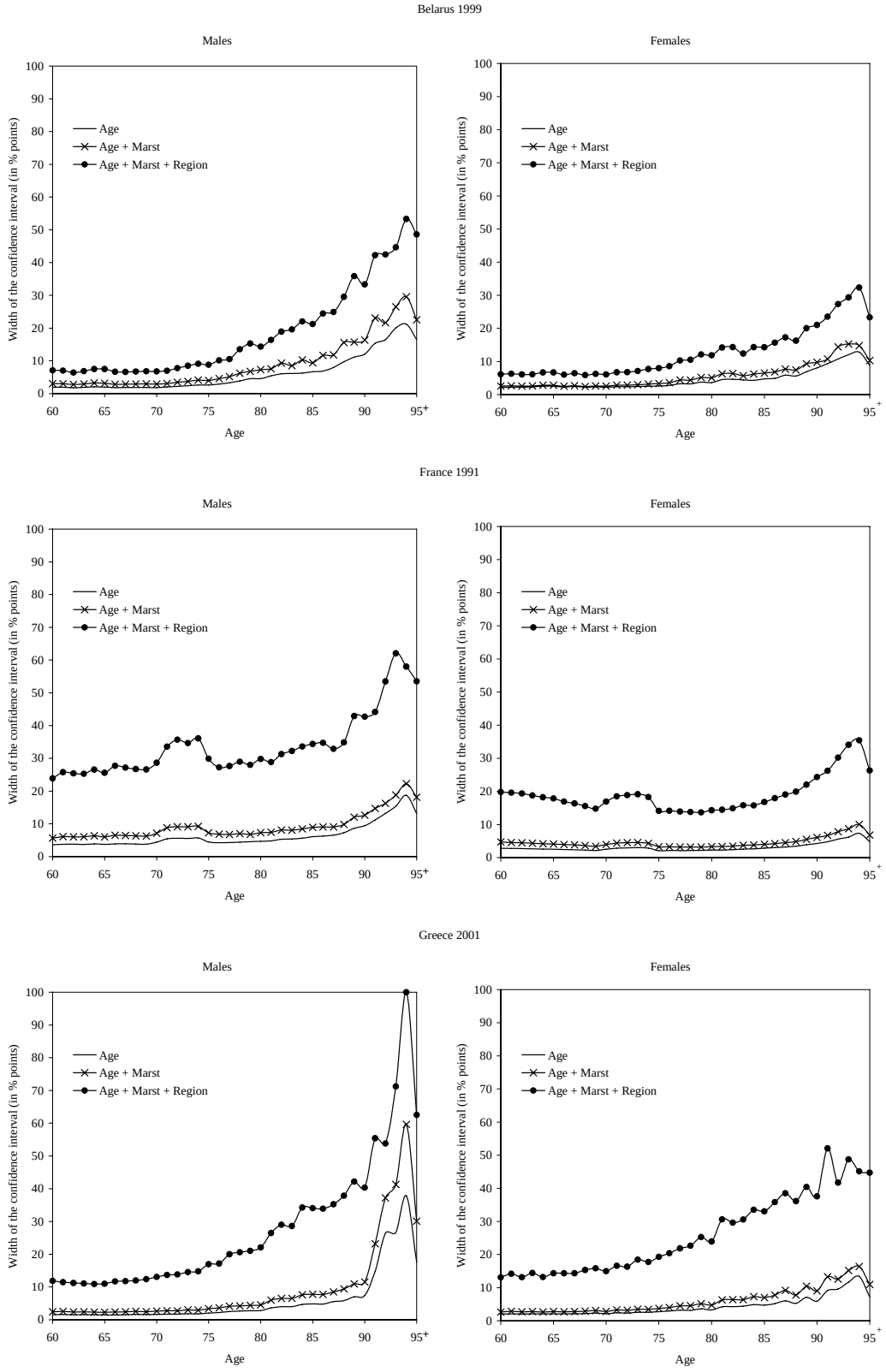


Figure 1 (continuation). Width of the confidence interval (in % points) of headship rates by age, marital status and region

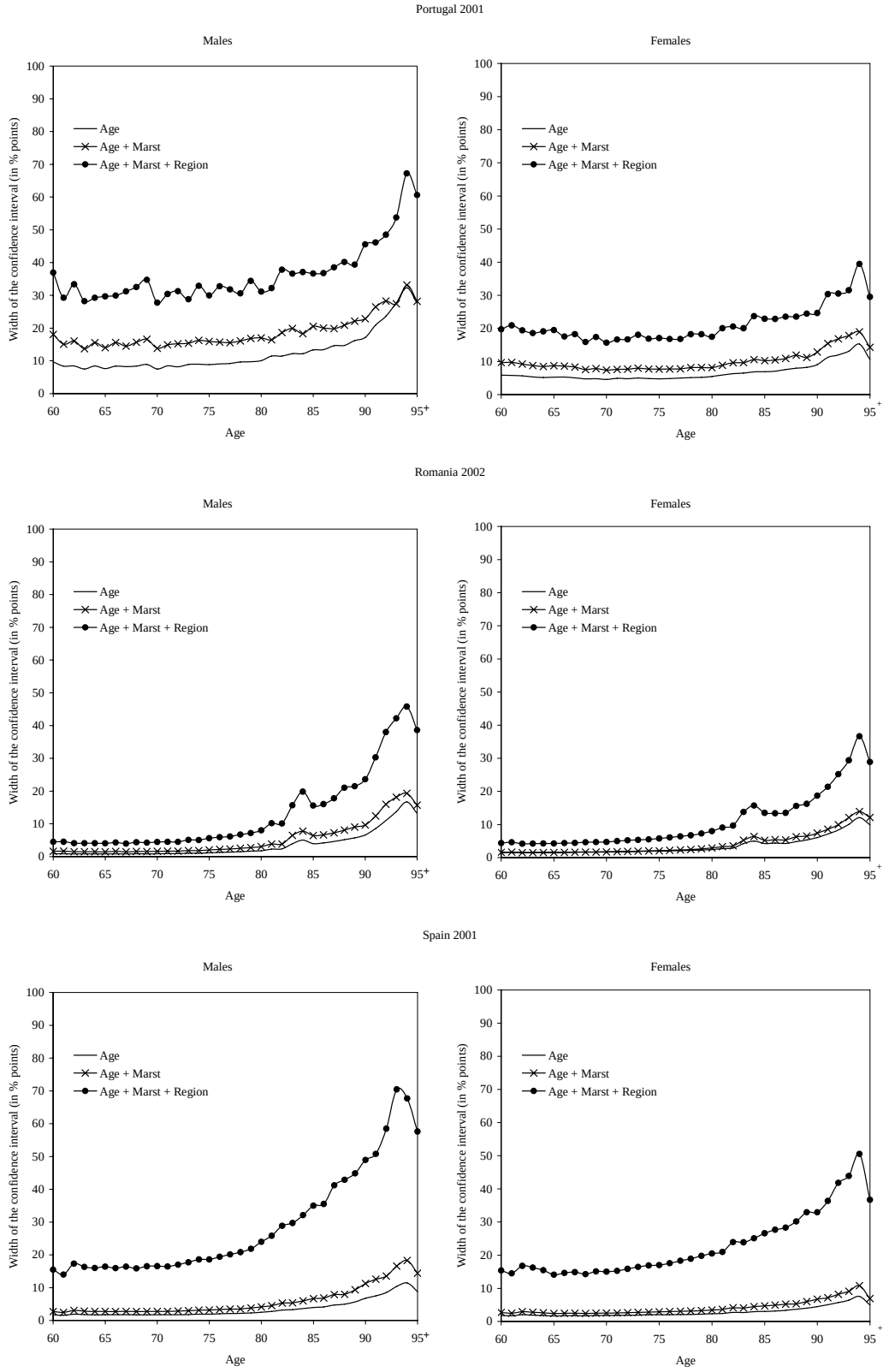


Figure 2. Width of the confidence interval (in % points) of headship rates by age, marital status and region according to hypothetical sample size increases

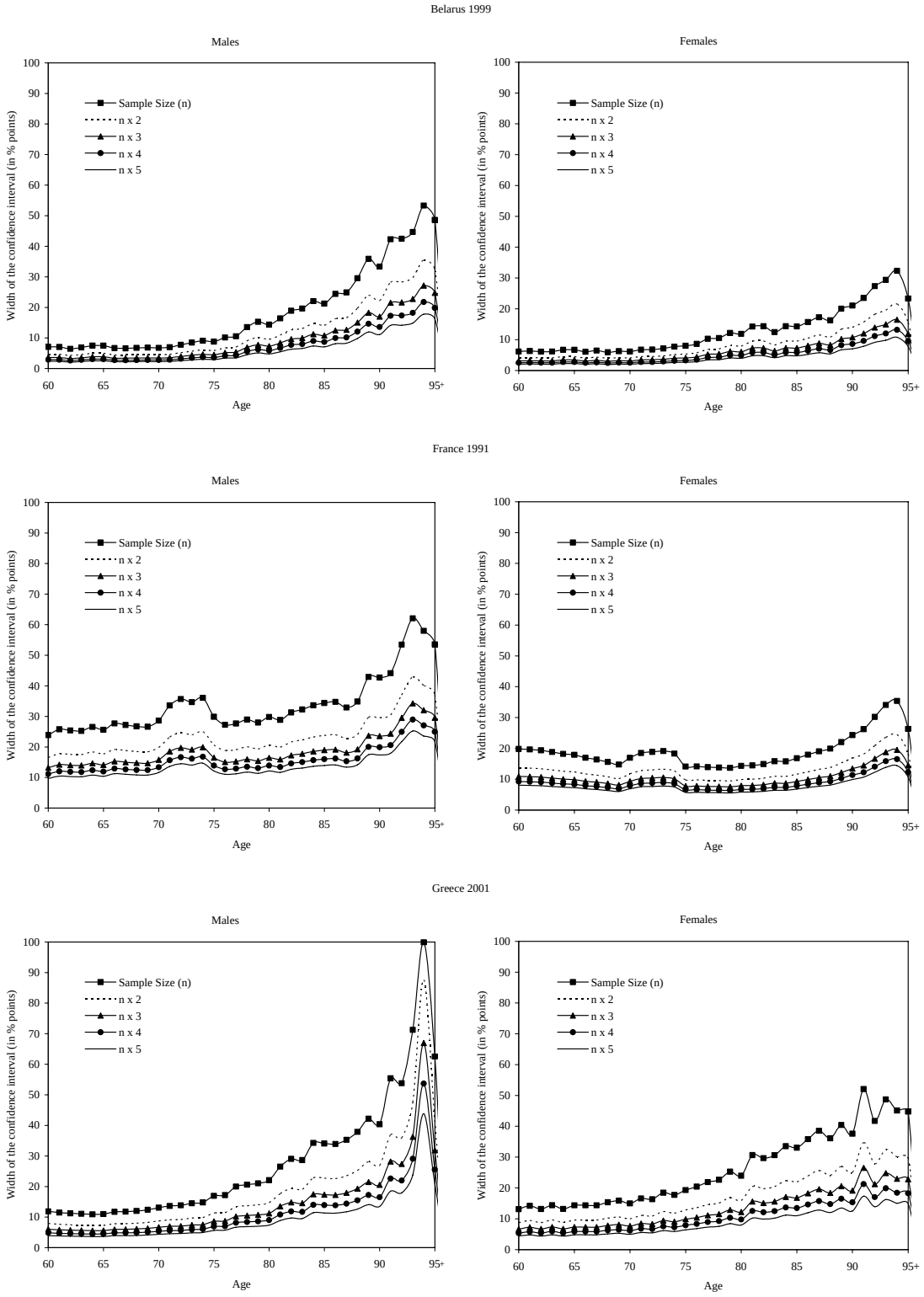
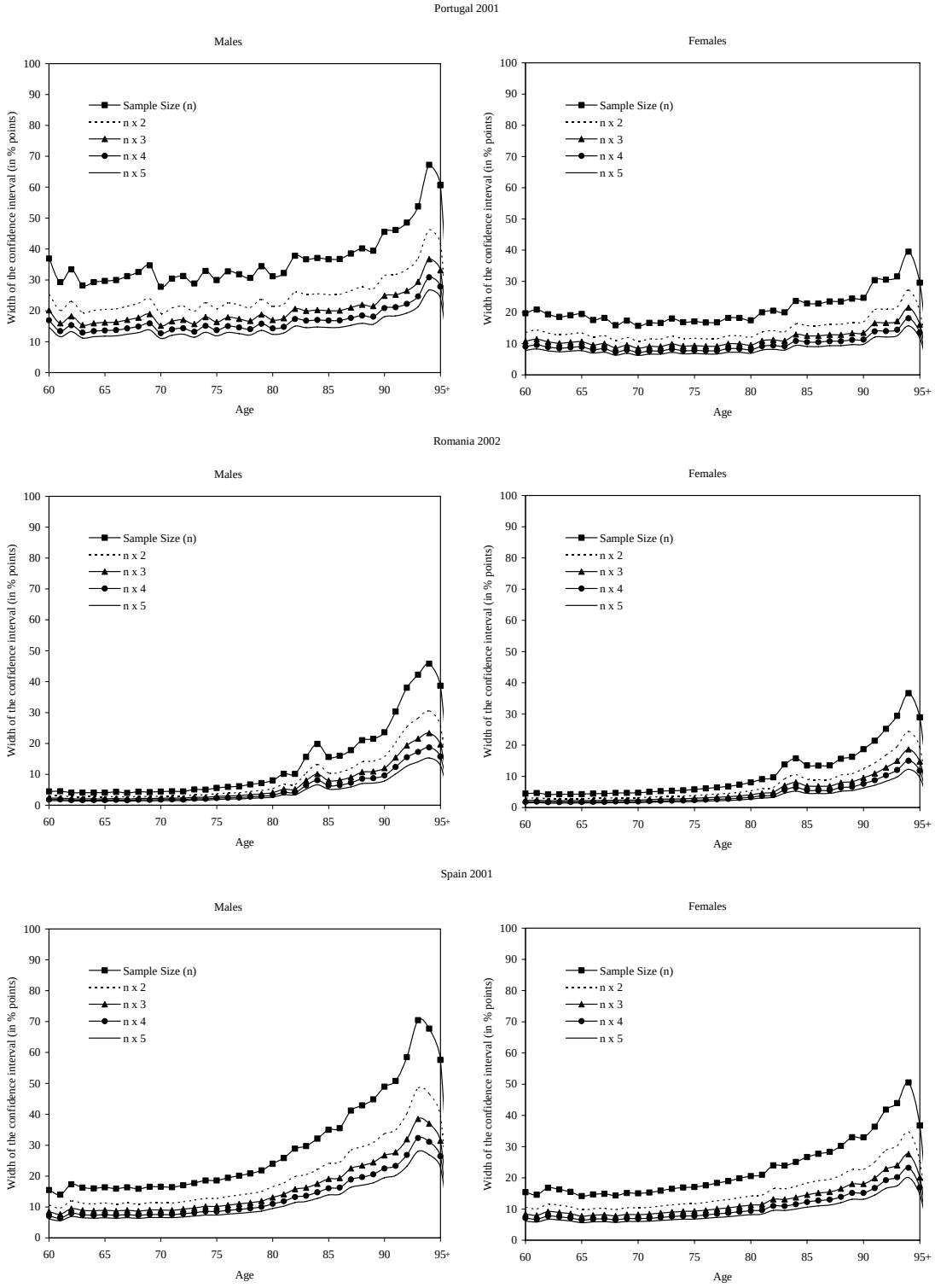


Figure 2 (Continuation). Width of the confidence interval (in % points) of headship rates by age, marital status and region according to hypothetical sample size increases



3 Conclusions

The IECM and IPUMS projects are fortunate to enjoy the support of many of the most respected National Statistical Offices in Europe. High precision household samples of 5-10% have been entrusted to the Barcelona-Minnesota team by the NSOs of eleven countries--Austria, Belarus, Czech Republic, France, Germany, Greece, Hungary, the Netherlands, Portugal, Romania, and Spain--and are expected soon from an additional four countries--Bulgaria, Italy, Switzerland, Turkey and the United Kingdom. While sample densities of 5-10% are adequate for most research purposes, this paper has shown that even higher densities—double, triple or even greater oversamples—are essential for the study of such important population sub-groups as the elderly. Otherwise sampling error alone will be so great as to deprive the results of significance. The Population Activities Unit (ECE), a precursor to the IECM project for the 1990 round of censuses, developed over-samples of up to 50% for Bulgaria, Czech Republic, Estonia, Finland, Hungary, Latvia, Lithuania, Romania, and Switzerland. In a second phase of the IECM initiative, we invite European NSOs to consider enriching the sample densities within a regimen of stringent controls for statistical confidentiality. By restricting access to a class of academic users, high-density microdata extracts can be provided to researchers at vanishingly low risk.

References

Boleslawsky L., E. Tabeau (2001), Comparing theoretical age patterns of mortality beyond the age of 80. In: E. Tabeau, A. Van den Berg Jeths, C. Heathcote (eds.), Forecasting mortality in developed countries: Insights from a statistical, demographic and epidemiological perspective, pp. 127-155. Kluwer Academic Publishers, Dordrecht.

Coleman, M., L. Ganong and M. Fine (2000). “Reinvestigating Remarriage: Another Decade of Progress.” Journal of Marriage and the Family 62(4): 1288-1307.

European Commission (2005), Network for the Integrated European Population Studies, NIEPS project final report, Brussels, EUR n° 21529, ISBN 92-79-00426-3.

Lopata, H. Z. (1996). Current Widowhood: Myths & Realities. Thousand Oaks, CA, USA: Sage.

McCaa, R., Esteve,A. (2006) 'IPUMS-Europe: Confidentiality measures for licensing and disseminating restricted-access census microdata extracts to academic users' in *Monographs of official statistics*, Eurostat, pp. 37-47.

Murray, C.J.L., A.D. Lopez (eds.), Global burden of disease and injury series: volume I: The global burden of disease: a comprehensive assessment and disability from diseases, injuries, and risk factors in 1990 and projected to 2020.

Olshansky S.J., Carnes B.A. & Cassel C. (1990) “In search of Methuselah: estimating the upper limits to human longevity” Science 250:634-640.

Olshansky S.J., Grant M., Brody J. and Carnes B. (2005) “Biodemographic perspectives for epidemiologists” Emerging Themes in Epidemiology 2(10) (http://www.ete-online.com/content/2/1/10).

United Nations (2007), World Population Prospects: The 2006 Revision. [data downloaded from http://esa.un.org/unpp on November 05, 2007].

United Nations (1997), Projecting old-age mortality and its consequences. Report on the Working Group, New York, 3-5 December 1996.

Yashin A. (2001), Mortality models incorporating theoretical concepts of ageing. In: E. Tabeau, A. Van den Berg Jeths, C. Heathcote (eds.), Forecasting mortality in developed countries: Insights from a statistical, demographic and epidemiological perspective, pp. 261-280. Kluwer Academic Publishers, Dordrecht.