IPUMS-International High Precision Population Census Microdata Samples: Balancing the Privacy-Quality Tradeoff by Means of Restricted Access Extracts

Robert McCaa, Steven Ruggles, Michael Davern, Tami Swenson,
and Krishna Mohan Palipudi

Minnesota Population Center, 50 Willey Hall
Minneapolis MN 55455 USA
contact: rmccaa@umn.edu

Abstract. A breakthrough in the tradeoff between privacy and data quality has been achieved for restricted access to population census microdata samples. The IPUMS-International website, as of June 2006, offers integrated microdata for 47 censuses, totaling more than 140 million person records, with 13 countries represented. Over the next four years, the global collaboratory led by the Minnesota Population Center, with major funding by the United States National Science Foundation and the National Institutes of Health, will disseminate samples for more than 100 additional censuses. The statistical authorities of more than 50 countries have already entrusted microdata to the project under a uniform memorandum of understanding which permits researchers to obtain custom extracts without charge and to analyze the microdata using their own hardware and software. This paper describes the disclosure control methods used by the IPUMS initiative to protect privacy and to provide access to high precision census microdata samples.

Keywords: Census microdata samples, data privacy, data quality, IPUMS-International

1 Introduction

In 1983, the legendary Charles M. Cawley offered the alumni association of his alma mater, Georgetown University, a deal. In exchange for its endorsement and a list of members, his fledgling credit card company, MNBA, would pay a percentage of revenues to the association. The offer was accepted and MNBA—by extending the affinity credit card offer to organizations with responsible, affluent members from the Association of Trial Lawyers of America to the Sierra Club—quickly established itself as the fastest growing, most profitable credit card company in the United States. Cawley became a billionaire. Now every successful credit card company in the world markets affinity cards.

The IPUMS project seeks neither profits nor popularity. Ours is a wholly academic initiative, but we target an affinity group, a “restricted class of individuals” [1] consisting of academic and policy researchers, who have great need to use population census microdata, but pose a vanishingly small risk of misuse.

Where much disclosure control research on the privacy-quality tradeoff is focused on either “public access” at one extreme or “safe-harbor” at the other [2], the IPUMS-International initiative adopts a third way, the “trusted user” approach [3]. Access is denied to approximately one-third of those who complete the electronic application form. Four years after dissemination began in May 2002, fewer than one thousand researchers have been granted access to IPUMS-International census microdata.

We restrict access to researchers who have a defined need to use the data and who not only agree to abide by the rigorous conditions of use license but also bind their institutions as enforcing agents. With, on the one hand, the assistance of our statistical agency partners, as stipulated in the project memorandum of understanding, and, on the other, the conditions of use license, misuse will lead to punishment not only for the individual but also for the individual’s institution. Indeed, in contrast to the record of commercial companies and government agencies, where there are frequent accounts of misuse of microdata for disclosing information about individuals, there is not a single, specific allegation of misuse of population census microdata in more than four decades of use by academic researchers. By rigorously policing access, we expect to extend this unblemished record of responsible scholarly use.

2 The Case for High Precision Samples: The USA Experience

In recent years, scholars working with United States census microdata have come to rely on high-precision samples. Beginning with the 1980 census, the Census Bureau has released five-percent samples as well as the one-percent samples. The five-percent samples for the United States in 1980, 1990, and 2000 include between 12 million and 14 million individuals in each year.

The Census Bureau anticipated that the 1980 five-percent sample would be used mainly for state and local policy analysis; at the time the sample was created, it was prohibitively expensive for most researchers to process the entire set of five-percent data. By the end of the 1980s, however, data processing costs had declined dramatically and were no longer a critical constraint for researchers at major institutions. Social scientists soon developed research strategies that capitalized on the availability of very large census microdata files. Swicegood et al. [4] published the first article in *Demography* that used a five-percent national sample, an analysis of language use and fertility in the Mexican-origin population. Later that year, Odland and Ellis [5] published a second *Demography* article using the large 1980 file, a study of household size and regional outmigration rates between 1975 and 1980.

From that time on, the use of high-precision census microdata files expanded rapidly. The cost of computing declined dramatically during the first half of the 1990s with the advent of inexpensive UNIX workstations. Moreover, during the past several years the performance of Windows-based desktop computers has improved to the point that a machine costing less than $1,000 is now easily capable of processing the five-percent samples of 1980, 1990 and 2000. Since 1996, the on-line data dissemination systems developed at Minnesota and elsewhere have provided easy access to large microdata extracts. Accordingly, the largest census microdata files—once available to few researchers at great expense—are now accessible, at no cost, to virtually all social scientists and policy analysts worldwide.

Increasingly, studies that use census microdata from 1980, 1990 or 2000 have turned to the five-percent files. Since 1990, 81 percent of *Demography* articles based on recent census microdata have used the high-precision samples.[1] Most of these analyses depend on information for small population subgroups, ranging from same-sex couples to the grandchildren of

[1] This percentage excludes eight articles that did not specify sample precision.

immigrants. In many instances, the large samples permit the use of innovative methods; to take just one example, these files have allowed demographers to carry out multi-level contextual analyses by making it feasible to assess the characteristics of small geographic areas.

The five-percent samples of the 1980, 1990 and 2000 censuses have now become the most widely used data source in the pages of *Demography*., as we learned from a analysis of the journal's pages in 2002. At that time, even though the United States had abundant high-quality survey data and the most recent census samples were over a decade old, high-precision census microdata files were used by a quarter of the articles on the United States that appeared in *Demography* in 2000 and 2001. In that period, the large samples were used twice as often as the next most popular data source. Clearly, the high-precision samples of the 1980 and 1990 censuses had become an indispensable component of American social science infrastructure. In 2003, with the addition of a five percent sample from the 2000 census, use skyrocketed.

It is impossible to determine an optimal size for a general-purpose sample. The number of cases needed to analyze a population subgroup depends on desired precision, type of subgroup, type of analysis, and population heterogeneity. If high precision estimates are required, many thousands of cases of the subgroup of interest may be necessary. Frequently, the relevant individuals for analysis are a small subset of the sample population. Multilevel analyses of the effects of local context on individual behavior are especially demanding since they often require data tabulated for small geographic units. The experience of the U.S. demonstrates that very large census microdata samples are among the most powerful tools available for economic and demographic analysis. As such samples become available for other countries around the world, they are becoming key components of social science and policy infrastructure.

3 The IPUMS Approach: High Precision Samples with Implicit Stratification

An important technique used to protect confidentiality of census microdata is to draw a high precision sample from all the census microdata records and then, in addition to the disclosure controls discussed below in sections 4 and 5, suppress from the sampled records all identifying information (names, addresses, and low-level geographical details). High precision samples preserve the ability to work with a large amount of microdata making it harder to identify any one person in the sample data file. In drawing high precision samples it is also important to think about efficient methods. By using stratification to draw a high precision sample, gains in efficiency are possible [6], [7]. To the extent the strata used to draw a high precision sample are associated with the variables of interest (e.g., orphanhood, poverty, unemployment, etc.), the resulting estimates of these variables will have lower standard errors than what would have resulted had a simple random sample of records been drawn from the complete census data [6], [7].

One of the most important stratifying variables in survey research and in drawing high precision census microdata samples is geography. Geography is related to a great number of variables researchers are interested in studying and therefore increases the efficiency of stratified samples. Many of the IPUMS-International samples capitalize on *implicit* geographic stratification. The raw census files used to create IPUMS samples are typically geographically organized within districts. Systematic random samples of the censuses capitalize on this low-level geographic sorting. By ensuring a representative geographic distribution of sampled cases, they are equivalent to extremely fine geographic stratification with proportional weighting. Since many economic and demographic characteristics are highly correlated with geographic location, this implicit stratification yields substantially greater precision than would a simple random sample of households. As part of the IPUMS project, we are developing stratification variables that allow researchers to make reliable variance estimates from implicitly stratified samples.

Almost all the statistical agency partners of the IPUMS project have endorsed the use of implicitly stratified samples of households (see Table 1, “sample design” column). Twenty-six countries (identified by “*” in Table 1) have provided complete sets of census microdata to facilitate the drawing of implicitly stratified samples by the project. In Europe, almost all the statistical agencies have drawn new samples using IPUMS specifications. IPUMS sample densities, as can be seen in Table 1, typically range between 5 and 10%. Lower densities are provided by countries where privacy matters are a greater issue than quality (Netherlands, United Kingdom) or, as in the case of 1960 round of censuses, where only low precision samples survive.

4 IPUMS-International Access Disclosure Controls

Access to the IPUMS-International database is governed, on the one hand, by the letter of understanding endorsed by the University and the National Statistical Authority, and, on the other by the license agreement between the University, the researcher, and the researcher's institution. Both are subject to amendment and enhancement as new methods are suggested. The letter of understanding grants the right to the university to disseminate microdata extracts electronically for teaching and research purposes via the project webpage: https://www.ipums.org/international, according to the authorization procedures stated in the agreement. Data may not be used for commercial purposes. Strict confidentiality of persons, households and other entities must be maintained. Alleging that a person or other entity has been identified is prohibited. The University is charged with assuring that users will guard against access to the microdata by unauthorized individuals.

The fact that IPUMS-International distributes microdata electronically as custom extracts, tailored as to country(ies), census year(s), subpopulation(s), and variables, according to the individual needs of the researcher, provides additional incentives for jealously guarding extracts. Since complete datasets are not distributed on CD or other medium, the inclination to share data with unauthorized individuals is greatly reduced, if not completely eliminated.

The electronic application form is designed to ascertain the bona fides of the applicant as well as the appropriateness of the microdata for the proposed research. A stern warning is issued against fraudulent applications, and checks are implemented to verify the identity and affiliations of the applicant (see the project home page "Apply for Access"). To confirm that the researcher understands the sensitivity of guarding the privacy of individuals, the application requests the name of the Human Subjects Protections Committee, Institutional Review Board, or similar office at the applicant's institution. A critical consideration in determining access is the proposed research. The statement must identify the data to be used and the purpose. Many applicants are denied access for failing to demonstrate that microdata are needed to address the proposed research or instructional plan. Finally the researcher must agree to seven restrictions on use: no redistribution, scholarly use only, prohibition on commercial use, strict rules of confidentiality, data security, appropriate citation, and notification of errors in the data. Approval is granted for a period of one year and may be renewed. Access to the microdata is password controlled. Remote data access is not offered. While this method might allow access to higher density, virgin microdata, our memorandum of understanding with the national statistical agencies does not authorize this form of access.

5 Technical Disclosure Controls

Where the statistical agency entrusts the anonymization procedures to the IPUMS project, we impose additional technical privacy protections. Technical controls are implemented on a subjective, ad-hoc basis as negotiated with each country for each census. Contemporary microdata, say from a census taken less than ten years ago, require more technical disclosure controls than older, historical data.

The most important technical control is the suppression of records by subsampling. All the values in the records outside the sample are suppressed. Second, is the suppression of names and geographical detail, such as place of birth or residence. Each statistical authority balances the trade-off by instructing the IPUMS project as to the minimum threshold for identifiable geographical units for the most recent census. In the case of many African and Latin American countries, the threshold is commonly set at 20,000 inhabitants in the latest census. Others place it as high as 100,000 (United States) or in the most extreme case (Netherlands) all administrative geography is suppressed. We are gratified that in some cases our statistical agency partners have reconsidered earlier decisions, offering higher precision samples (Mexico 1990 increased from one to ten percent) and greater detail. In the case of Colombia, the geographical threshold, initially set at 100,000, was reduced to 20,000 after Colombian geographers vigorously registered their dissatisfaction. The Colombian statistical agency not only reduced the threshold, but also harmonized the identifiers so that all the census microdata samples for Colombia could be disseminated with a single set of geographical codes.

Additional protection is provided by randomly ordering the records and swapping the geographical identifiers of an undisclosed number of households. This means that no one can state with certainty that an individual or household has been identified.

In consultation with the national statistical office, some variables may be top-coded, others may be subjected to global recoding, deletion of digits for hierarchical variables (occupation, industry, geography), or the suppression of a variable entirely. Decisions are made in consultation with the corresponding national statistical authority. Sensitive variables, if any, may be suppressed entirely at the request of the statistical agency. Weight variables are usually not an issue because most of the samples are implicitly stratified with a single weight. We do not resort to either microaggregation or Post Randomization (PRAM) methods.

6 Countering Fear, Hysteria and Paranoia with Reason

Privacy rights and statistical confidentiality of data are severely threatened by government, commercial firms, and individuals—but the threat to population census microdata is virtually nil. Fear, hysteria and paranoia are incited among official statisticians by the widespread circulation of a "pizza commercial" developed by an American civil liberties advocacy group [8] and advertisements offering private details of individuals and entities for a price. What is striking is that none involve population census microdata. Indeed, there is no market—black, grey, gold or otherwise—for anonymized census microdata samples for the purpose of identifying individuals or linking to other data sources. Even in the United States, at a moment of shocking violations of individual rights by government agencies, there is not one allegation of access to census microdata by the Homeland Security Agency or other government agencies. The reason is obvious. Population census microdata samples, per se, do not contain sensitive or valuable political or commercial information, and without personal identifiers, statistical linkage is useless due to the high proportion of false positives [9].

7 Conclusion

The goal of IPUMS is to restore balance to the privacy-quality tradeoff by providing high precision, anonymized samples to a restricted class of researchers. In the IPUMS datasets identification is impossible for the vast majority of persons and positive identification is always impossible. Given the wealth of information readily available from private sources in most countries, it would be foolhardy to turn to census microdata to attempt to uncover imprecise and outdated information about a particular individual. We invite academics who need census microdata for research purposes to examine the offerings at the IPUMS website.

Acknowledgements. Funded in part by the National Science Foundation of the United States, Grant No. SES0433654.

References

1. Willenborg, L., de Waal, T.: Elements of Disclosure Control. New York: Springer-Verlag (2001)
2. Willenborg, L., de Waal, T.: Statistical Disclosure Control in Practice. New York: Springer-Verlag (1996)
3. McCaa, R., Esteve, A.: IPUMS-Europe: Confidentiality measures for licensing and disseminating restricted-access census microdata extracts to academic users. In Monographs of official statistics: Work session on statistical data confidentiality. Luxembourg: Office for Official Publications of the European Communities, (2006) 37-46.
4. Swicegood, G., Bean, F.D., Stephen, E.H., Opitzm, W.: Language Usage and Fertility in the Mexican-Origin Population of the United States. Demography. 25 (1988) 17–33
5. Odland, J., Ellis, M.: Household Organization and the Interregional Variation of Out-migration Rates. Demography. 25 (1988) 567-579
6. Kish, L.: Weighting for Unequal P_i. Journal of Official Statistics. 8 (1992) 183-200
7. Kish, L.: Survey Sampling, Wiley Classics Library Edition. New York: Wiley and Sons (1995)
8. American Civil Liberties Union (ACLU). Surveillance Campaign. (2005) Available online at http://www.aclu.org/pizza/
9. Dale, A., Elliot, M.: Proposals for 2001 SARS: An assessment of disclosure risk. Journal of the Royal Statistical Society. Series A. 164, part 3 (2001) 427-447

Table 1. IPUMS-International: 160 microdatasets entrusted by country, subsample precision and design
For current data availability, see: https://www.ipums.org/international

Datasets entrusted by subsample precision									
10%	**~5%**	**<=4%**	**Country**	**Sub sample design**	**2000s**	**1990s**	**1980s**	**1970s**	**1960s**
Release 1, May 2003 (28 datasets)									
5			**Brazil**	IPUMS	**2001**	**1991**	**1980**	**1970**	**1960**
		1	**China (only ‘82 ‘til now)**		**2000**	**1990**	**1982**		1964
3		1	***Colombia**	IPUMS		**1993**	**1985**	**1973**	**1964**
	5		**France (’99 in preparation)**	IPUMS	**1999**	**1990**	**1982**	**1975**	**1968, 2**
	2		**Kenya (’79 & ‘69 in process)**	IPUMS	**1999**	**1989**	**1979**	**1969**	
2		2	**Mexico (’80 in recovery)**	IPUMS	**2000**	**1990**	1980	**1970**	**1960**
	5		**United States**		**2000**	**1990**	**1980**	**1970**	**1960**
	2		**Vietnam**	IPUMS		**1999**	**1989**	1979	
Release 2, June 2006 (19 datasets)									
4		1	***Chile**	IPUMS	**2002**	**1992**	**1982**	**1970**	**1960**
3	1		***Costa Rica**	IPUMS	**2000**		**1984**	**1973**	**1963**
4		1	***Ecuador**	IPUMS	**2001**	**1990**	**1982**	**1974**	**1962**
2			**South Africa**		**2001**	**1996, 1**	**1985, 0**	**1970**	1960
3			***Venezuela**	IPUMS	**2001**	**1990**	**1981**	**1971**	**1961**
Europe (27 datasets)									
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
			Germany (in process)		**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 in recovery)**	IPUMS	**2001**	**1992**		**1977**	1965
			Russia (negotiating)		**2002**		**1989**	1979	1970
			Slovenia		**2001**	**1991**	**1981**		

Table 1. IPUMS-International: 160 microdatasets entrusted by country, subsample precision and design
For current data availability, see: https://www.ipums.org/international

Datasets entrusted by subsample precision				**Sub sample design**					
10%	**~5%**	**<=4%**	**Country**		**2000s**	**1990s**	**1980s**	**1970s**	**1960s**
	3		**Spain**	IPUMS	**2001**	**1991**	**1981**	1970	1960
			Switzerland (negotiating)		**2000**	**1990**	**1980**	**1970**	1960
			Turkey (in process)		**2000**	**1990**	1980, 5	**1970, 5**	1960, 5
		1	**United Kingdom (in process)**		**2001**	**1991**	**1981**	**1971**	**1961**
North America and the Caribbean (27 datasets)									
		3	**Canada**		**2001**	**1991,** 6	**1981,** 6	**1971,** 6	1961, 6
1	1	2	***Dominican Republic**	IPUMS	**2003**	**1993**	**1981**	**1970**	**1960**
1			***El Salvador**	IPUMS		**1992**		1971	1961
2		3	***Guatemala**	IPUMS	**2002**	**1994**	**1981**	**1973**	**1964**
3		1	***Honduras**	IPUMS	**2000**		**1988**	**1974**	**1961**
1			***Nicaragua**	IPUMS	**2005**	**1995**		**1971**	1963
5			***Panama**	IPUMS	**2000**	**1990**	**1980**	**1970**	**1960**
	4		**Puerto Rico**		**2000**	**1990**	**1980**	**1970**	1960
South America (17 datasets)									
4			**Argentina**	IPUMS	**2001**	**1991**	**1980**	**1970**	1960
3			***Bolivia**	IPUMS	**2001**	**1992**		**1976**	
4		1	***Paraguay**	IPUMS	**2002**	**1992**	**1982**	**1972**	**1962**
1			***Peru**	IPUMS		**1993**	**1981**	1972	1961
4			***Uruguay**	IPUMS		**1996**	**1985**	**1975**	**1963**
Africa (17 datasets)									
2			**Egypt**	IPUMS		**1996**	**1986**	1976	1964
2			***Guinea, Conakry**	IPUMS		**1996**	**1983**		1960
			Lesotho (in process)			**1996**	**1986**	**1976**	1966
1			***Madagascar**	IPUMS		**1993**			
2			***Malawi**	IPUMS		**1997**	**1987**	**1977**	1967
3			***Mali**	IPUMS		**1998**	**1987**	**1976**	
2			***Rwanda**	IPUMS	**2002**	**1991**			
3			***Sudan**	IPUMS		**1993**	**1983**	**1973**	
2			***Uganda**	IPUMS	**2002**	**1991**	1980		1969
Asia and Oceania (25 datasets)									
1			**Armenia**	IPUMS	**2001**		1989	1979	1970

Table 1. IPUMS-International: 160 microdatasets entrusted by country, subsample precision and design
For current data availability, see: https://www.ipums.org/international

Datasets entrusted by subsample precision				**Sub sample design**					
10%	**~5%**	**<=4%**	**Country**		**2000s**	**1990s**	**1980s**	**1970s**	**1960s**
			Bangladesh (in process)		**2001**	**1991**	**1981**	1974	1961
1			**Cambodia**	IPUMS		**1998**			1962
3			***Fiji Islands**	IPUMS		**1996**	**1986**	1976	**1966**
			Indonesia (in process)		**2000**	**1990**	**1980**	**1971**	1961
1			***Iraq**	IPUMS		**1997**	1987	1977	1967
4			**Israel**	IPUMS		**1995**	**1983**	**1972**	**1961**,7
		4	**Malaysia**		**2000**	**1991**	**1980**	**1970**	1960
1			***Mongolia**	IPUMS	**2000**		1989	1979	1970
3			***Pakistan**	IPUMS		**1998**	**1981**	**1973**	1961
1			**Palestinian Authority**	IPUMS		**1997**			
3		2	***Philippines**	IPUMS	**2000**	**1990**	**1980**	**1970**	**1960**
1			**Turkmenistan**	IPUMS		**1995**	1989	1979	1970

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