When Excessive Perturbation Goes Wrong and Why IPUMS-International Relies Instead on Sampling, Suppression, Swapping, and Other Minimally Harmful Methods to Protect Privacy of Census Microdata

Lara Cleveland, Robert McCaa, Steven Ruggles, and Matthew Sobek

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

Abstract. IPUMS-International disseminates population census microdata at no cost for 69 countries. Currently, a series of 212 samples totaling almost a half billion person records are available to researchers. Registration is required for researchers to gain access to the microdata. Statistics from Google Analytics show that IPUMS-International's lengthy, probing registration form is an effective deterrent for unqualified applicants. To protect data privacy, we rely principally on sampling, suppression of geographic detail, swapping of records across geographic boundaries, and other minimally harmful methods such as top and bottom coding. We do *not* use excessively perturbative methods. A recent case of perturbation gone wrong—the household samples of the 2000 census of the USA (PUMS), the 2003–2006 American Community Survey, and the 2004–2009 Current Population Survey—, an empirical study of the impact of perturbation on the usability of UK census microdata—the Individual SARs of the 1991 census of the UK—, and a mathematical demonstration in a timely compendium of statistical confidentiality practices confirm the wisdom of IPUMS microdata management protocols and statistical disclosure controls.

Keywords: population census, microdata samples, data privacy, data dissemination, statistical disclosure controls, IPUMS-International

1 Introduction

IPUMS-International is a global collaboratory of universities, national statistical authorities, data repositories, and research centers to archive, integrate, and disseminate census microdata [1], [2]. Founded in 1999 and led by the Minnesota Population Center, the project currently disseminates 212 confidentialized, integrated population census

samples, representing 69 countries and totaling almost one-half billion person records. Each year the database is updated with samples for the latest 2010 round censuses and for five to ten additional countries as integration of their microdata is completed. For the 2010 census round (2005-2014), samples for eighteen countries are already integrated into the database (Table 1). In 2013, we expect to add 2010 round samples for an additional ten countries: Brazil, Burkina Faso, Cameroun, Fiji Islands, Ghana, Israel, Kenya, Kyrgyz Republic, Panama, and the USA plus an additional sixteen samples for earlier censuses. At current growth rates, by the end of the decade, census samples for more than 100 countries are likely to become available through the IPUMS-International portal.

Table 1. 69 countries with integrated population microdata available June 2012 from www.ipums.org/international (number of samples in parentheses)

Africa 28 samples	*Egypt (2), Ghana (1), Guinea (2), Kenya (2), *Malawi (3), Mali (2), Morocco (3), Rwanda (2), Senegal (2), Sierra Leone (1), *South Africa (3), *Sudan (1—includes South Sudan), Tanzania (2), Uganda (2)
Americas 81samples	Argentina (4), Bolivia (3), Brazil (5), Canada (4), Chile (5), *Colombia (5), Costa Rica (4), Cuba (1), Ecuador (5), *El Salvador (2), Jamaica (3), *Mexico (7), *Nicaragua (3), Panama (5), *Peru (2), Puerto Rico (5), Saint Lucia (2), *USA (7), *Uruguay (5), Venezuela (4)
Asia and Oceania 47 samples	*Cambodia (2), China (2), India (5), *Indonesia (9), *Iran (1), Iraq (1), Israel (3), Jordan (1), Kyrgyz Republic (1), Malaysia (4), Mongolia (2), Nepal (1), Pakistan (3), *Palestine (2), Philippines (3), Thailand (4), *Vietnam (3).
Europe 56 samples	Armenia (1), Austria (4), Belarus (1), *France (7), Germany (4—includes GDR and FRG), Greece (4), Hungary (4), *Ireland (8), Italy (1), the Netherlands (3), Portugal (3), Romania (3), Slovenia (1), Spain (3), Switzerland (4), Turkey (3), the United Kingdom (2)

Note: * = sample for a 2010 round population census is already integrated into the IPUMS-International database

2 Restricted Access, Customized Census Microdata Extracts

Access to the IPUMS-International microdata is free of cost, but restricted. Despite the "PU" in IPUMS, the microdata are not "public use".[1] Would-be users must submit a detailed electronic application both to establish research bona-fides and to explain need for access. An essential part of the process is to agree, individually, to ten stringent restrictions on condition of use—prohibiting redistribution, restricting to scholarly use, prohibiting commercial usage, protecting confidentiality, assuring security, enforcing

[1] The full moniker is "Integrated Public Use Microdata Series".

strict rules of confidentiality, permitting scholarly publication, citing properly, threatening disciplinary action for violations, and reporting errors.

Google Analytics suggest that the IPUMS-International registration form alone is a substantial deterrent to unqualified users. Over a recent twelve month period, 5,593 views of the registration page yielded only 1,057 completed applications. A significant reason for the large drop-off is that the registration form poses a daunting deterrent to the statistically naive. First, eighteen bits of personal and professional information must be entered into the form. Second, the applicant must identify the name of the Human Subjects Protection Committee of his or her institution, a supervisor's name and email address, a website listing the individual's institutional affiliation, and telephone number. Third, the applicant must agree to abide by each of ten restrictions on usage noted above. Fourth, a project description (75 words minimum) must be entered into a text box on the form. Finally, the applicant should select the countries for which microdata are desired. Optionally, the applicant may also indicate countries of interest for which microdata are currently unavailable from the IPUMS-I website.

A qualified user with a genuine research need will readily fill-out the application and provide the requested information in meticulous detail, regardless of time required to complete the registration form. The unqualified, on the other hand, will not complete the form at all. Incomplete forms are automatically rejected by IPUMS web page controls. It is impossible to submit an incomplete application. The daunting detail required to complete the form leads to self denial by all but the highly motivated researcher.

Once the registration is submitted, applicants are carefully vetted to prevent access by both the unqualified and those who lack a research need. In calendar year 2011, a mere 46, of the 1,057 completed applications were denied access as a result of the vetting process. The most frequent reason for denial is that the currently disseminated census microdata are not suitable for the proposed research. The second most frequent reason is that no microdata for the country requested are currently available in the database. A few would-be users are denied access because the database lacks a crucial variable needed for the research (e.g., current real estate value of dwelling).

For qualified researchers, the registration form educates users to guard the microdata against misuse. Over the past decade, more than 5,000 users world-wide—representing almost a thousand institutions and over one hundred nationalities—have successfully registered and in doing so have bound themselves and their institutions to stringent terms of use. More than one-third of IPUMS-I's trusted users request access to microdata for a single country. Many of these are resident abroad who seek access to data for their country of identity.

From this brief description of the IPUMS-International registration process it should be apparent that before an individual account is activated, we do due diligence to confirm the identity and research bona fides of each applicant. IPUMS-International is not simply a click-and-get website. Nonetheless, we firmly believe that delay is the deadliest form of denial. We strive to complete the review process within a day or two and at most a week.

Agreeing to the conditions of use binds both the researcher and the researcher's institution. The Legal Counsel of the University of Minnesota is poised to strike at the

first indication of misuse. A violation by a single user will suspend access to all users at that institution, until researchers undergo remedial training for the protection of human subjects and the institution regains its accreditation for handling sensitive microdata.

Thanks to these procedures and others, IPUMS-International is the only academic organization disseminating international census microdata that is cited as good practice by the Conference of European Statisticians Task Force on Managing Statistical Confidentiality and Microdata Access [3].

IPUMS-International distributes microdata electronically as custom extracts, tailored with regard to country(s), census year(s), subpopulation(s), and variables, according to the individual needs of the researcher. Microdata may be requested for multiple countries and census years. For each request, the microdata are pooled into a single file. Moreover IPUMS-International offers powerful value-addeds, such as the "attach characteristics" and "select cases" features. The average extract in calendar year 2011 consisted of a mere 35 variables, including six technical variables that are automatically included with each request. Dissemination of such highly customized microdata provides additional incentives for users to jealously guard their extracts. Since complete datasets are not distributed on CDs or any other media, the temptation to share microdata with unauthorized individuals is greatly reduced.

The IPUMS-International method of disseminating extracts contrasts with the practices of most official statistical agencies, which deliver microdata as a product, often a labeled CD or DVD. Typically, under this old-fashioned approach, when requests are fulfilled, each researcher receives exactly the same set of documentation and microdata sample containing all variables and all person records. Given the massive size of the IPUMS-International database, disseminating the full set of variables and unvarying size of samples is simply impossible.

3 Sampling, Suppression and Swapping Efficiently Protects Data Privacy and Statistical Confidentiality

The microdata disseminated by IPUMS-International are subjected to strong, uniform legal and administrative controls, providing greater protections for all participating statistical agencies as a group than for any single office that chooses to go it alone[1], [2]. Technical disclosure control standards consist of two types: IPUMS-International standards and national [2, see table 1]. National standards are usually idiosyncratic and almost invariably undocumented.

All apply two of the most powerful privacy protection controls: first, the suppression of names and low level geographical detail; second, the suppression of records by the use of sub-sampling. All the values in the records outside the sample are excluded. For group quarters, communal establishments, large households and such, we shift the unit from households, and sample only individuals.

In addition, with respect to IPUMS-International standards, each statistical authority balances the confidentiality/utility trade-off by instructing the Minnesota Population Center as to the minimum threshold for identifiable geographical sub-units. For many countries, the threshold is commonly set at 20,000 inhabitants. Others place it as high as 100,000 (United States) or in the most extreme case (the Netherlands) all administrative geography is suppressed.

In consultation with the national statistical office, we top-code some variables, global-recode others, selectively delete digits of those with hierarchical codes (occupation, industry, geography), or even suppress variables entirely. Decisions are made in consultation with the corresponding national statistical authority. In the case of the 2001 census of Switzerland, 68 categorical variables had one or more code suppressed, and 12 continuous variables were top-and-bottom coded. Household with more than 15 persons were considered "group quarters" and sampled as individuals.

Additional powerful statistical disclosure protections are provided by randomly ordering the records and swapping the lowest-level geographical identifiers of an undisclosed number of paired households [4, 171]. Swapping on geographic attributes is an exceedingly strong method for assuring confidentiality at minimal loss of data utility. Swapping across geographical boundaries means that no statement that an individual or household has been identified can be made with absolute certainty. After an exhaustive review of statistical disclosure control methods, the Registrar Generals of England and Wales, Scotland and Northern Ireland adopted swapping of low-level geographical attributes as the principal method for protecting confidentiality for both the microdata and the tabular outputs of the 2011 census of the UK[5]. The loss of data utility is limited to the level of geography at which swapping occurs[6]. At higher levels the harm to the data is nil, yet confidentiality of the microdata is protected by the fact that drilling down to the lowest level of geography may yield a persons and household that in fact has been swapped in from a different geographical location. Allegations of identification may be made, but the uncertainty remains regarding the true location of the individual.

4. When Excessive Perturbation Goes Wrong

Perturbation for the purpose of statistical disclosure control adds noise to the value of a data attribute [5, see pp. 112-114]. IPUMS-International does not use excessive perturbative methods, but two of our partners did for the 2000 round of censuses: the United States Census Bureau and the Office of National Statistics (UK). Unfortunately, mistakes were made in the case of the 2000 census microdata of the USA[2] (including the five percent household sample disseminated by both IPUMS-International and IPUMS-USA). Alexander et. al. pointed out that there were substantial discrepancies between the

[2] Similar errors were discovered in the microdata files of the 2003–2006 American Community Survey and the 2004–2009 Current Population Survey [6].

number of men and women at each individual age between the published counts and the microdata [7]. After stories about the botched anonymization were published in the New York Times [8], Washington Post, and Wall Street Journal, the Census Bureau re-released the microdata with "corrected" age information.[3] The corrections addressed the specific statistical discrepancy cited by Alexander et al., but it did not fix the problem.

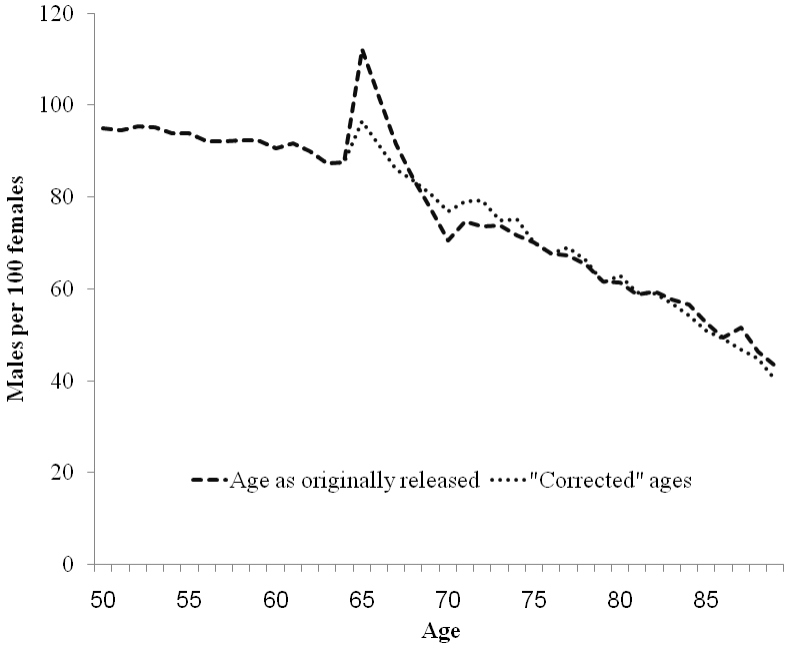

Fig. 1. Sex ratio in Census 2000 Microdata: original and "corrected" ages.

Figure 1 compares the sex ratio using the excessively perturbed microdata originally released by the Census Bureau with the sex ratio calculated from the "corrected" data. The original data overstated the sex ratio at age 65 by almost 30%. The corrected data is quite a bit better, but still overstates the sex ratio at age 65 by more than 10%.

Depending on the measure, the corrected microdata are sometimes even more distorted than the original microdata. Most of the relationships between age and other characteristics measured for individuals are reasonable, but the new perturbation method does not account for the characteristics of family members. While the addition of noise is designed to preserve means and covariances, it is impossible to retain all possible relationships in a hierarchically structured microdata file [9]. Household is the sample

[3] IPUMS-USA microdata files were updated Aug. 10, 2011. "Age" contains the corrected data. "Ageorig" is as originally released by the Census Bureau. The IPUMS-International file was updated with the 2012 release.

unit for all IPUMS-USA and almost all IPUMS-I datasets, including the 2000 census of the USA. Many researchers use these high precision samples to analyze the relationships between members of households, such as the age difference between spouses.

Figure 2 compares the mean age difference between husbands and wives (calculated as age of husband minus age of wife). We compare the measure as calculated from the age as originally released and the corrected ages. In addition, we compare with the 2001 American Community Survey, which was not excessively perturbed and therefore probably reflects the true pattern of spouse intervals.

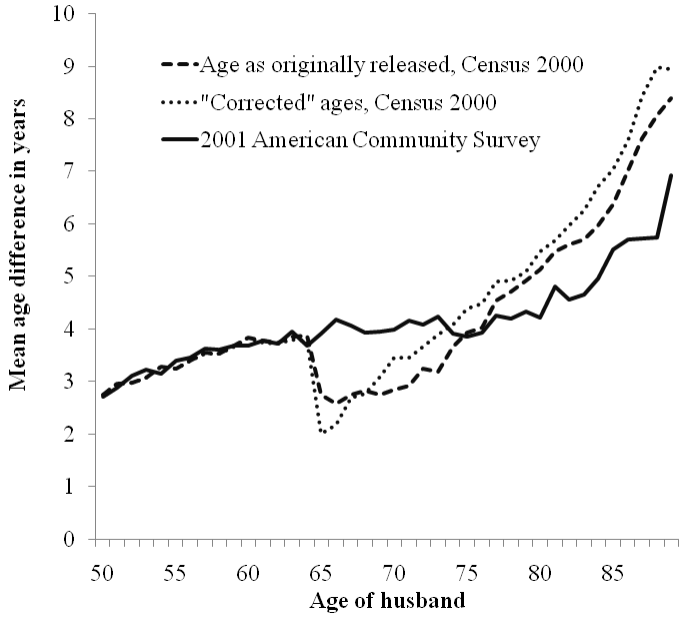

Fig. 2. Mean difference between husband's age minus wife's, Census 2000 and 2001 American Community Survey.

The excessively perturbed data understates spouse intervals from age 65 until the mid 70s, and overstates the intervals from the late 70s to the late 80s.

At most ages, the problem is significantly worse for the corrected data than it is for the original perturbed data. For example, at age 65 the new data understates the spouse interval by approximately 50%, but the original version of age understated it by only 30%. At age 87, the corrected data overstate the interval by 47 percent, whereas the original data only overstated it by 33 percent. Either error is of an unacceptable magnitude, but the point is that even when the problem was explicitly pointed out, the leading Census Bureau perturbation experts got it wrong again. Unless you know just how the data are going to

be used, it is really impossible to design a perturbation method that does not introduce new, harmful sources of error.

Purdam and Elliot reached a similar assessment after conducting an exhaustive replication of ten published studies based on UK Samples of Anonymized Records (SARs) using both perturbed and unperturbed microdata. They conclude: "our study does indicate that the perturbations applied by the μ-Argus system have a significant impact on the outcome of analyses" [9, p. 1111]. For the general case, Duncan, Elliot and Salazar-Gonzalez offer a mathematical demonstration that "correlation coefficients in the masked data will appear closer to zero than they are in the source data" [5, p. 113].

We suspect that synthetic microdata is likewise unsuccessful in replicating analytical interrelationships inherent in high precision household population census samples. Three years ago as a test case, we entrusted—with permission of the official census agency-owner—a full count microdataset to a major figure in the synthetic microdata field with the challenge of constructing a dataset. The test would be a real-world, side-by-side comparison of substantive conclusions, comparing differences in specific models between the original and synthetic data. Unfortunately, the test ended without results because no synthetic microdataset was produced. In the words of the analyst, an expert with 10 synthetic microdata publications cited by Google Scholar, "research has not caught up yet with the complexity and dimensionality of such real-life datasets".[4] Nonetheless the "black-box" challenge to analyze the comparative substantive utility of synthetic versus real microdata remains on the table.

5. Conclusion.

In May 2002, with the launch of the IPUMS-International website and the first release of twenty-one census samples encompassing six countries, we have relied on sampling, suppression and swapping to confidentialize microdata and facilitate access to the database by researchers around the globe—with restrictions, but without cost. We adopted these procedures because of their proven effectiveness in protecting privacy and statistical confidentiality with minimal loss of data utility. A decade later, the recent, unfortunate US Census Bureau's experience with excessive perturbation and the embrace of swapping by the Office of National Statistics of the UK confirm the wisdom of our procedures. Soon we expect to launch two new modes of access. First will be a basic tabulator operating behind the current password protected system. For researchers needing simple frequencies and two or three-way cross-tabulations, this will alleviate the need to submit and download a full-scale extract. Second will be a remote access system, the IPUMS-International Remote Data Center, where researchers may analyze higher sample densities—in some cases as high as 100%—and more detailed geographies—as low as municipalities, localities (NUTS4 or 5, in the EuroStat scheme) and even

[4] Private email dated, May 20, 2012.

enumeration districts—without actually downloading the microdata. Readers interested in contributing to either of these innovations are invited to contact the authors.

Acknowledgements. Funded in part by the National Science Foundation of the United States, Grant Nos. SES-0433654 and 0851414; National Institutes of Health, Grant Nos. R01HD047283 and R01 HD044154.

References

1. McCaa, R., Ruggles, S., Davern, M., Swenson, T., and Mohan Palipudi, K. IPUMS-International High Precision Population Census Microdata Samples: Balancing the Privacy-Quality Tradeoff by Means of Restricted Access Extracts. In Domingo-Ferrer, J. and Franconi, L. (eds.). Privacy in Statistical Databases. PSD2006 Proceedings, LNCS 4302 Berlin: Springer-Verlag (2006) 375-382.
2. McCaa, R., Ruggles, S., and Sobek, M. IPUMS-International Statistical Disclosure Controls: 159 Census Microdata Samples In Dissemination, 100+ In Preparation. In Domingo-Ferrer, J. and Magkos, E. (Eds.). Privacy in Statistical Databases. PSD2010 Proceedings, LNCS 6344. Heidelberg: Springer (2010) 74-84. DOI: 10.1007/978-3-642-15838-4_7
3. United Nations Economic Commission for Europe. Conference of European Statisticians. Managing Statistical Confidentiality & Microdata Access: Principles and Guidelines of Good Practice. Geneva: United Nations (2007) See online edition Annex 1.23: http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf
4. Reiter, J.P. Statistical Approaches to Protecting Confidentiality for Microdata and Their Effects on the Quality of Statistical Inferences. Public Opinion Quarterly (2012) 76(1): 163-181. doi:10.1093/poq/nfr058
5. Duncan, G.T., Elliot, M., and Salazar-González, J.-J. Statistical Confidentiality: Principles and Practice. Heidelberg: Springer (2011).
6. Frend, J., Abrahams, C., Forbes, A., Groom, P., Spicer, K., Tudor, C., and Youens, P. Statistical Disclosure Control in the 2011 UK Census: Swapping Certainty for Safety. ESSnet Workshop on Statistical Disclosure Control (SDC) of Census Data, Luxembourg, April 19-20, 2012.
7. Alexander, J.T., Davern, M., and Stevenson, B. Inaccurate Age and Sex Data in the [United States] Census PUMS Files: Evidence and Implications. Public Opinion Quarterly (2010) 74(3): 551-569. doi:10.1093/poq/nfq033
8. Wolfers, J. Can You Trust Census Data? Freakonomics blog. New York Times, February 2, 2010. http://freakonomics.blogs.nytimes.com/2010/02/02/can-you-trust-census-data
9. Purdam, K. and Elliot, M. A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymised Records. Environment and Planning A (2007) 39(5): 1101-1118.