ISSN 1727-9917

SERIES

STUDIES AND PERSPECTIVES

ECLAC SUBREGIONAL HEADQUARTERS FOR THE CARIBBEAN

Dissemination of Caribbean census microdata to researchers

Including an experiment in the anonymization of microdata for Grenada and Trinidad and Tobago

Francis Jones
Kristin Fox

UNITED NATIONS

ECLAC

SERIES

STUDIES AND PERSPECTIVES

ECLAC SUBREGIONAL HEADQUARTERS FOR THE CARIBBEAN

49

Dissemination of Caribbean census microdata to researchers

Including an experiment in the anonymization of microdata for Grenada and Trinidad and Tobago

Francis Jones
Kristin Fox

UNITED NATIONS

ECLAC

This document has been prepared by Francis Jones, Population Affairs Officer, of the Statistics and Social Development Unit of the subregional headquarters for the Caribbean of the Economic Commission for Latin America and the Caribbean (ECLAC), and Kristin Fox, consultant (formerly Databank Manager, the Derek Gordon Databank, University of the West Indies).

The views expressed in this document, which has been reproduced without formal editing, are those of the authors and do not necessarily reflect the views of the Organization.

United Nations publication
ISSN 1727-9917
LC/L.4134
LC/CAR/L.486
Copyright © United Nations, February 2016. All rights reserved.
Printed at United Nations, Santiago, Chile
S.15-01381

Member States and their governmental institutions may reproduce this work without prior authorization, but are requested to mention the source and inform the United Nations of such reproduction.

Contents

Tables

Figures

Abstract

Caribbean census microdata are not easily accessible to researchers. Although there are well-established and commonly used procedures technical, administrative and legal which are used to disseminate anonymized census microdata to researchers, they have not been widely used in the Caribbean. The small size of Caribbean countries makes anonymization relatively more difficult and standard methods are not always directly applicable. This study reviews commonly used methods of disseminating census microdata and considers their applicability to the Caribbean. It demonstrates the application of statistical disclosure control methods using the census datasets of Grenada and Trinidad and Tobago and considers various possible designs of microdata release file in terms of disclosure risk and utility to researchers. It then considers how various forms of microdata dissemination: public use files, licensed use files, remote data access and secure data laboratories could be used to disseminate census microdata. It concludes that there is scope for a substantial expansion of access to Caribbean census microdata and that through collaboration with international organisations and data archives, this can be achieved with relatively little burden on statistical offices.

Introduction

Over the last twenty five years, statistical offices worldwide have increasingly sought to meet the demand from researchers for greater access to statistical microdata. Statistical disclosure control methods have been developed to enable statisticians to release microdata in a controlled way which protects the privacy and statistical confidentiality of individuals and other entities. There has been a substantial growth in the volume of microdata made available to researchers in universities and other organizations. Census microdata is among the most useful to social researchers because of the range of social and demographic information collected in censuses, the information which is available about small population subgroups, and the historical comparability of censuses.

Census microdata are the raw data about individuals and households as opposed to the aggregated statistics appearing in a published census report. Statistical disclosure control methods make it possible for statistical offices to anonymize census microdata so that there is a low risk of individuals and households being identified within the data. Such methods make it possible to disseminate census microdata to researchers in universities or in government thus more fully exploiting its potential value for social research and policy analysis.

The confidential nature of the information provided by households is guaranteed in national laws governing census taking and is also one of the United Nations Fundamental Principles of Official Statistics (United Nations, 2014). Information is collected on the understanding that it will be treated as confidential and census respondents are generally given some kind of guarantee or assurance in this regard. Official statisticians are charged with guarding the confidentiality of this information and the respondent s trust in the statistical office s adherence to this commitment is an important factor in their willingness to participate in the census.

Census micro-level datasets are tremendously rich sources of information and the scope for analysing them to answer different research questions is considerable. Dissemination of microdata facilitates more sophisticated analysis than is possible based on tables of aggregated data. Allowing researchers from outside the statistical office to access this data dramatically increases the number of analysts who are able to study the data thus making much greater use of this valuable statistical resource. Where a national statistical office goes to the huge expense of running a census, publishing the results in a traditional census report, but without subsequently allowing analysts from other organizations to

access the microdata, a valuable resource is being wasted. A statistical office can only ever hope to carry out a small fraction of the analysis that could potentially be of great value in social research and policymaking. The task of the official statistician is to design procedures and select methods which strike a balance between these conflicting imperatives: granting some form of access for researchers to census microdata while still protecting the confidentiality of the information provided by individual census respondents.

The dissemination of census microdata can be thought of as one of three levels of dissemination of census results. The first level consists of traditional census reports and census tables published in print and online. These census outputs are aimed at the widest audience of census users.

The second level of dissemination is through interactive tabulation tools (for example using ECLAC s REDATAM software). This provides users with the capacity to generate their own census tables (or graphs or maps) by submitting queries to a database containing the census microdata but without allowing the user direct access to the microdata records themselves. Interactive tabulation facilities are generally provided online and are aimed at census users possessing data analysis skills.

The dissemination of census microdata would be the third level of this schema. Datasets consisting of anonymized record level data for individuals and households are provided to researchers to enable them to carry out the kinds of analyses and research that are only possible with direct access to microdata, for example analyses which involve statistical models of social and demographic phenomena. Users of census microdata are generally skilled researchers. To protect the confidentiality of individual census records, census microdata disseminated in this way are anonymized through the use of statistical disclosure control methods.

Caribbean countries and territories generally publish traditional census reports. Some have developed interactive tabulation tools using REDATAM (Aruba, Belize, Saint Lucia, Saint Vincent and the Grenadines and Trinidad and Tobago). A similar number have used statistical disclosure control methods to anonymize census records for researchers (Belize, Jamaica, Saint Lucia, Suriname and the United States Virgin Islands). However, Caribbean census microdata remain an underused resource which hinders social research and policy development.

Caribbean countries suffer from a paucity of statistical sources making it doubly important to fully exploit those that do exist. Release of census microdata would help to maximise the use of the data, avoid possible duplication of effort in data collection activities, and improve the harmonization and comparability of scientific research. Furthermore, feedback from researchers can help to improve the reliability of the data and force producers to pay greater attention to quality.

There is general agreement on the value and on the need to make microdata available to researchers. The Standing Committee of Caribbean Statisticians (SCCS), in 2010, adopted the following position on access to microdata:

- Access to all statistical data, whether microdata or tabular data, shall be strictly in accordance with the Statistics Acts of CARICOM Member States and Associate Members;
- Member States and Associate Members are encouraged to establish mechanisms for disclosure prevention, such as data anonymization and to provide access to microdata under controlled conditions, such as microdata laboratories;
- International organisations are encouraged to build capacity at the national level and support the establishment of these mechanisms in CARICOM Member States and Associate Members.

The issue, therefore, is not whether to disseminate microdata but rather how to disseminate it in a manner that will maximise its utility to researchers while preserving the confidentiality of record-level data. A combination of technical, administrative, and legal controls can be used to strike this balance.

The technical procedures are the statistical disclosure control methods which are most commonly applied in the design of microdata release files. These methods include the removal of direct identifying variables, sampling of records, recoding of variables, suppression of variables or categories within

variables, and methods of data perturbation such as data swapping, data shuffling or post randomisation. A more indirect form of microdata release is to provide remote access facilities (RAF) where the data is held on a secure server but researchers can submit queries, written in a particular programming language, with the results returned online. In this case, the statistical disclosure control methods have to be embedded into the RAF infrastructure.

The administrative procedures refer to the application processes, the data access agreements signed by researchers in order to gain access to microdata, and the procedures for monitoring and dealing with breaches of the terms of those agreements. Data access agreements place obligations upon researchers and/or their institutions to protect confidentiality and not to seek to identify individuals or households in the dataset. The knowledge that any breach of those terms by the researcher or institution could lead to future denial of access to microdata or legal action, is a powerful incentive to adhere to the conditions of the data access agreement. An alternative to a signed agreement is to ask researchers to agree to certain terms and conditions of access in order to obtain the data, for example by clicking I agree in order to download a file. This is clearly a weaker form of protection since there is no control over who downloads the file.

It is considered good practice for data access agreements to have legal force to increase public confidence that microdata will be used appropriately (United Nations, 2007). However, this requires a proper legal framework and statistical legislation, or other regulation, which addresses the dissemination of microdata and provides for legally enforceable data access agreements. In Caribbean countries, such legal frameworks are not generally in place. For example, statistics acts make no specific reference to the release of microdata. However, while a legislative framework for the release of microdata and legally enforceable agreements are preferable, they are not essential since there is nothing in law which prevents the release of microdata to researchers provided that the technical and administrative arrangements protecting confidentiality are sufficiently strong.

Additional factors to be taken into account when considering the appropriate arrangements for the release of microdata are the researcher or organization that will receive the data, the nature of their research and the use to which the data will be put. Microdata are generally released for scientific research, rather than for commercial purposes, and so for these reasons are mainly disseminated to universities, research institutes and public sector organizations.

For Caribbean countries, disclosure control is especially challenging because of their small population size. Disseminating analytically useful microdata, while still protecting confidentiality, is relatively more difficult for censuses carried out in small countries where identification of individuals or households from their census responses is more feasible. Consequently, many Caribbean statistical offices have taken a very cautious approach to dissemination.

The Statistical Institute of Belize (SIB) is one of the countries that has advanced furthest in the provision of microdata having put in place a Microdata Access Program over the last two years. The SIB uses disclosure control methods to release anonymized census microdata (and other microdata) to researchers who sign a confidentiality agreement. There is also a secure data laboratory which researchers can visit, on-site, to analyse more complete micro-level datasets under the supervision of the Data Dissemination Department. Staff of this department carry out disclosure checks and have to approve all outputs before they are allowed to leave the premises. The General Bureau of Statistics of Suriname also offers similar microdata laboratory facilities.

The Statistical Institute of Jamaica disseminates census data to researchers through the Derek Gordon Databank based at the Mona campus of the University of the West Indies in Kingston, Jamaica. Researchers have to apply for access to the data. They are required to submit an abstract or other description of their proposed research and sign a confidentiality agreement in which they undertake to adhere to stringent conditions concerning the use of the data. Researchers are provided with a subset of the microdata which is designed to meet their research needs while protecting statistical confidentiality. The Derek Gordon Databank makes available data from the population censuses of Jamaica for 1982, 1991 and 2001. The 2010 data have been lodged with the databank but have yet to be distributed to researchers.

The statistical offices of Jamaica and Saint Lucia also make census microdata available through the IPUMS International project (Integrated Public Use Micro-data Series), the world s largest international archive of census microdata run by the Minnesota Population Center (MPC) at the University of Minnesota. Similarly, researchers must apply to access the data, and the MPC employ technical, administrative and legal procedures to protect confidentiality. Microdata from the Jamaican censuses of 1982, 1991, and 2001 are available as are data from the Saint Lucian censuses of 1991 and 1980. Data from the 2010 round of censuses has not yet been made available. Trinidad and Tobago has also been collaborating with the MPC and it is anticipated that Trinidad and Tobago census microdata will soon be available through the IPUMS International project. In addition to anonymization, the integration and harmonisation of census data is an important part of the IPUMS International project.

The purpose of this study is to analyse how Caribbean census microdata can be made more widely available to researchers. It considers how international practices can be adopted and adapted to the Caribbean; carries out a detailed anonymization of census data from Grenada and Trinidad and Tobago; and drawing on the lessons from this exercise, it proposes how access to Caribbean census data can be expanded.

I. Dissemination of Census Microdata

Statistical offices and data archives disseminate census microdata to researchers using a range of technical, administrative and legal procedures to protect the privacy of individuals and households, and the confidentiality of the information they provide to census takers. This chapter discusses these techniques and procedures in more detail and draws some general conclusions about their applicability to the Caribbean.

A. Technical disclosure control methods for census data

1. Disclosure control concepts and scenarios

When statistical confidentiality is breached, it is referred to as disclosure. Disclosure is considered to have occurred when someone is able to discover new information about specific individuals, households (or other entities) from published statistics. In the context of the release of census microdata, this occurs when a user of the microdata is able to verify, with an unacceptably high degree of probability, that a record (or records) in the microdata file belongs to a particular individual or household. This is referred to as identity disclosure or re-identification. Having established and verified a link between a record in the census microdata and an individual, the user then has access to the full census record for that person.

There is another form of disclosure referred to as attribute disclosure which is theoretically possible. This would arise if a user was able to discover information about an individual, not by linking them to a single census record, but by deducing that their census record must be one of a small group of records. However in the context of census microdata release, this is very much a secondary concern. Attribute disclosure is more difficult to establish and yields much less information. Therefore in this analysis, the risk of disclosure is considered to be the risk of identity disclosure.

Analyses of disclosure risk posit the existence of users that seek to re-identify individuals and households in the released microdata. These hypothetical users are referred to as intruders or attackers. They may be motivated by a wish to find out information about one or many individuals for either personal reasons, commercial reasons, or simply by a desire to discredit the statistical office.

There are two main disclosure scenarios relevant to the dissemination of census microdata. One is often referred to as the external archive scenario . In this case it is assumed that the intruder attempts to

match the release file with some other dataset, register, or database of administrative records which contain similar key variables (for example age, sex, marital status, place of birth etc.) Based upon the information in these key variables they attempt to re-identify individuals in the microdata release file. In this case, the intruder is attempting to re-identify a large number of individuals, in a systematic way, but is limited by whatever information their external database contains.

The other scenario is the so-called nosy neighbour scenario. This supposes that an intruder seeks to find the census records belonging to people among their circle of acquaintances based on the information they know about them. A nosy neighbour might only threaten to disclose the identity of a very small number of census records. However, it is reasonable to assume that a sufficiently motivated nosy neighbour could amass a significant amount of information about an individual, and particularly a whole household, and would therefore have at least some chance of being able to re-identify an individual.

Analyses of disclosure risk commonly distinguish between different kinds of variables in a micro-level dataset.

- Direct identifying variables are those which include information such as names, addresses and telephone numbers. These variables are always removed from the dataset; in most cases they are not useful for statistical analysis.
- Indirect or quasi-identifying variables are those which could be used to re-identify at least some individuals or households. Unlike direct identifiers they are important variables for statistical analysis and so should be retained in the dataset as far as possible (Domingo-Ferrer and Torra, 2008). Examples include age, sex, place of residence, marital status, ethnicity, religion, occupation, industry and place of birth.
- Non-identifying variables are those which are not considered to be particularly useful for an intruder trying to systematically re-identify individuals. This could be because the information contained in these variables is not available in other datasets, registers or databases; because the information is not likely to be in the public domain; or because the variables do not partition the population in a way which poses a risk of re-identification. However, these variables may contain information which would be considered sensitive if it were to be disclosed, for example information about health. Where they measure rare population characteristics, these variables could be used by a nosy neighbour to re-identify someone, so it is necessary to protect against this.

The likely sensitivity of census information is an additional factor to be taken into consideration. Either indirect-identifying or non-identifying variables could be sensitive depending on the social context. Information about health conditions, disabilities or religious beliefs are generally considered personal information and disclosure of such information could lead to discrimination against individuals. For example, the Trinidad and Tobago 2011 Census asked whether household members suffered from HIV/AIDS (with some doubts surrounding the accuracy of the responses). Bearing in mind the stigma and discrimination against persons with HIV, this clearly constitutes sensitive personal information.

It should be emphasised that statistical laws do not generally make any distinction between sensitive and non-sensitive information. All information collected through the census is deemed confidential whether it is sensitive or not. Nevertheless, the likely sensitivity of the information is relevant to the extent that it could influence both the likelihood of an attack and the seriousness of the consequences of identity disclosure. Therefore sensitive variables, particularly variables related to health conditions, should be treated with extra care.

2. Analysis of disclosure risk and methods of disclosure control

Micro-level datasets are anonymized using methods of statistical disclosure control. It is important to note that it is generally very difficult to reduce the risk of disclosure to zero. The only certain way to achieve this is not to release any microdata. Methods are chosen with a view to reducing the risk of disclosure to an acceptable level while seeking to preserve, as far as possible, the utility of the dataset to

the researcher. Striking a balance between disclosure risk and data utility for different categories of user is the essence of statistical disclosure control.

There are various approaches to analysing disclosure risk but all depend on assumptions about the key variables . These are the indirect identifying variables: age, sex, place of residence, marital status, occupation, place of birth etc.; information which it is supposed that an intruder possesses and uses in an attempt to re-identify individuals. It is assumed that an intruder attempts to re-identify individuals rather than households since individual level variables contain more information which an intruder could use for this purpose. A household would be regarded as having been re-identified when at least one individual in the household has been re-identified.

One approach to analysing disclosure risk is to simulate a matching exercise between a proposed microdata release file and some other dataset which is thought to be indicative of the information that an intruder may have available to them. This approach depends on being able to identify a plausible external database with which to conduct a matching exercise.

An alternative approach is to analyse the rareness or uniqueness of records with respect to key variables, in both the microdata release file and the population as a whole. Since records which are rare or unique are at greater risk of being re-identified by an intruder, it is possible to draw conclusions about disclosure risk based on the number of such cases. In particular, it is records which are unique in the population which are at the greatest risk since an intruder that can establish the uniqueness of an individual in the population will be able to uniquely identify their record in a microdata release file (assuming the record for that individual is in the file). Using this idea, measures of the risk of re-identification can be calculated for individual records. These measures of the riskiness of individual records can then be aggregated to provide global measures of risk applying to the whole file.

Templ (2008) argues that the most effective way of masking microdata is by doing the anonymization steps in an exploratory and in some sense iterative way. It is possible to apply various methods on various variables with different parameters resulting in different effects on the data while looking for sufficient anonymization of the data with respect to low information loss and low re-identification risk.

There are two broad classes of methods for disclosure control: methods of data reduction and perturbation methods. Methods of data reduction, as the name suggests, reduce in some way the information that is released to researchers. Perturbation methods on the other hand introduce small adjustments into the data in order to hinder the efforts of intruders to link records to individuals. In particular, they make it possible to counteract claims that specific individuals have been re-identified with certainty.

a) Data reduction methods

As part of the anonymization process, some variables are simply removed from the dataset. This normally applies to: direct identifying variables; low level geographic variables; variables relating to aspects of census administration which are not relevant to researchers; variables where the data is of poor quality; and variables which cannot be protected by recoding.

Variables are often recoded to reduce the level of detail in the dataset thereby reducing the risk of disclosure. Categories are combined or, in the case of continuous variables, cases are grouped into intervals. Examples of this global recoding of variables would include coding age into five-year bands, or reducing the level of detail in a hierarchical classification, for example from six digit occupation or industry codes to three or even two digit.

Ordinal or continuous variables are often top or bottom coded. This means that all the cases above, or below, a certain threshold are grouped into a single category, for example all ages above 90 might be coded into a single category. This prevents identity disclosure for individuals that have uncommonly high, or low, values on those variables.

One of the most important methods of data reduction applied to census microdata is sampling (that is, the suppression of non-sampled records). Sampling reduces the risk of re-identification because

it reduces the number of records released and because an intruder cannot presume that records which may be unique in the sample are necessarily unique in the population. A disadvantage of sampling is that estimates and analyses produced from a sample of records will differ from those that would have been produced using the full census dataset. This problem is particularly acute for the smaller Caribbean census datasets where a sample of records will be of marginal utility to researchers simply due to its small size. For example, while a 10 per cent sample of 1 million person records would yield 100,000 records for analysis, a 10 per cent sample from 100,000 would yield just 10,000 records for analysis. A small sample of records will make it particularly difficult to analyse relatively rare population characteristics or events such as health conditions, disabilities or births and deaths.

b) Data perturbation methods

The other common class of disclosure control methods are data perturbation methods. These techniques modify variables in the original dataset to hinder the efforts of intruders to re-identify individuals. The perturbations introduced may be random or deterministic but should not be so significant that they excessively distort the data or would influence the analysis of researchers.

The most common perturbative method applied to census microdata is data swapping which has been used by the US Census Bureau, the United Kingdom s Office for National Statistics and in the IPUMS International project. It is normally applied to geographic codes. It involves taking a small percentage of households and for each of these identifying a matching household of similar basic characteristics and then swapping the geographic codes between these households, literally swapping the location of the households. Matching pairs of households are normally identified within the same area, at a certain level of the country s administrative geography. This means that the data swapping has no effect at higher levels of geographic aggregation. For most analyses it is unlikely to introduce major distortions of the microdata. One of the reasons that statistical offices prefer data swapping is that it is easy to implement as well as the fact that marginal distributions are preserved exactly at higher geographic aggregations of the data (Shlomo, Tudor and Groom, 2010).

A range of other data perturbation methods have been proposed including the Post Randomisation Method (PRAM) and data shuffling. PRAM randomly reassigns the values of a categorical variable according to transition probabilities which control the degree of perturbation. Another perturbative procedure called data shuffling, which can be applied to categorical, ordinal and continuous variables, has also been proposed for use on census data (McCaa and others, 2013).

However, bearing in mind the statistical capacity constraints in the Caribbean subregion, it would be more pragmatic to focus attention on tried and trusted methods of disclosure control whose use for census microdata is well-established, rather than on more complex and potentially risky perturbation procedures. Indeed, there is proven evidence that complex perturbation procedures, if not applied with great care and attention, can easily introduce distortions and inconsistencies which damage the data to an extent which either prevents its use in research, or even worse, affects research outcomes (see Cleveland and others, 2012).

It is very common to use methods of data reduction in tandem with perturbation methods. This is because methods of data reduction alone will tend to destroy the utility of the data before they have reduced disclosure risk to an acceptable level. Therefore methods of data reduction are applied first, up to a certain point where they have reduced disclosure risk significantly, but before they have done excessive damage to data utility. At that point, the presence of some remaining risky records in the release file is accepted, and then a method of data perturbation is applied in order to protect these records and introduce uncertainty into any attempt to re-identify these individuals.

B. Administrative arrangements for access to census data

Just as important as the technical controls to protect statistical confidentiality, are the policies, administrative procedures and agreements governing access to microdata: the categories of researchers to whom the datasets are released; how they are granted access; the conditions under which they are granted access; and the obligations that they assume to protect confidentiality.

There are three commonly used modes of microdata release, all of which can be used for the release of census microdata: release of public use files (PUF); release of licensed use files; and provision of controlled access to microdata in secure data laboratories (see for example Dupriez and Boyko, 2010).

1. Public use files

PUFs are readily available to the public under a set of basic conditions, including that they are not to be sold, shared, used for commercial purposes, linked with other datasets or used to identify individuals. Strong technical disclosure controls are applied so that the risk of re-identification of individuals is minimal. Public use files have two very appealing characteristics. From the point of view of the user they are very easy to obtain, generally via download. From the point of view of the statistical office, public use files have the advantage that once they have been published, there is no ongoing requirement to manage requests for the data.

There is one small Caribbean territory which is currently disseminating census microdata in the form of a public use file. That is the United States Virgin Islands (with a population of approximately 105,000). The United States Virgin Islands has the advantage that its census is carried out as part of the United States Census Bureau s operation. It therefore benefits from that organizations accumulated expertise in statistical disclosure control methods. The US Census Bureau in particular tends to favour public use rather than licensed use microdata release due to the legal interpretation of the Census Bureau s statute: that the data either are public or they are not, and that if they are public then they must be made available to any user (United Nations, 2007).

The microdata release file for the United States Virgin Islands has been produced using similar methods to those used to produce the United States Public Use Microdata Samples (PUMS), that is, a systematic sample of households with data provided for all individuals in selected households, global recoding, top and bottom coding, a minimum threshold for the number of cases in categorical variables, and some special treatment for persons living large households and the institutional population (US Census Bureau, 2015). The application of these methods has to be tailored somewhat to the small population size of the United States Virgin Islands. The PUF for the United States Virgin Islands is produced using a 10 per cent sample of households, with inclusion of all individuals in each household, single year of age (top coded at 88), with all geography codes removed (US Census Bureau, 2013). The file is available for anyone to download from the US Census Bureau website.

The example of the PUF for the United States Virgin Islands is interesting. It has been argued that the release of PUFs might not be possible in small countries due to the greater ease with which individuals could be re-identified. The United Nations Economic Commission for Europe s guide to Managing Statistical Confidentiality and Microdata Access , produced in 2007, argued that the level of disclosure risk will be much greater for countries with smaller populations, and therefore researchers should not expect that all countries will be able to release PUFs (United Nations, 2007). However, the United States Census Bureau, which has been publishing Public Use Microdata Samples since 1970, and has perhaps more experience of releasing PUFs than any other organization in the world, has determined that it is able to release a PUF for a territory with a population of 105,000.

What is possible in one country is not automatically possible in another country because disclosure risk depends not only on population size but also on the environment into which the data is released, for example the existence and availability of other datasets which could plausibly be matched with a census microdata file. Nevertheless, the availability of the PUF for the United States Virgin Islands is instructive, and suggests that the dissemination of samples of anonymized census microdata for small countries is at least technically possible. Historically, statistical organisations with a legal responsibility to guarantee statistical confidentiality have tended to adopt an overly conservative approach to the release of microdata and have over-estimated the risk of disclosure. The idea that PUFs cannot be made available in small countries may also turn out to have been an overly conservative assumption.

2. Licensed use files

Licensed use files are used to provide researchers with more detailed microdata which cannot be disseminated in the form of PUFs. Disclosure control methods are still applied to licensed use files but the protection is somewhat lighter than for PUFs. In particular, licensed files could be disclosive if they are matched against other datasets or registers. One of the main purposes of the licensing arrangements is to prevent such data linking.

Researchers must apply for access to licensed use files and are normally expected to provide a description of the proposed research. They must meet certain eligibility criteria concerning their organisation and their research to ensure that the microdata are used for statistical and scientific research. In addition, they must sign a data access agreement which governs their use of the data and obliges them to ensure that there is no breach of statistical confidentiality. These agreements may cover individual researchers, research teams or entire organisations.

Licensing of files provides significant protection against the risk of disclosure because it restricts access to persons who have a genuine need to access the data for research purposes. For researchers, there are substantial disincentives to break the terms of a data access agreement including the threat of legal proceedings depending on national laws. Institutional and professional sanctions are also a strong disincentive, for example denial of future access to microdata would lead to substantial reputational damage to a researcher and their institution (McCaa, Ruggles and Sobek, 2010). The census datasets disseminated by the Statistical Institute of Belize, the Statistical Institute of Jamaica, and the Central Statistical Office of Saint Lucia are all disseminated to researchers under license conditions.

Licensed use files also tend to be samples of census records. Sampling of records provides very strong protection of statistical confidentiality. Even if a researcher breaks the terms of a data access agreement and matches a sample of census records with an external database, they would still face very great difficulties in identifying individuals and households with any degree of certainty. The datasets disseminated through the IPUMS-International Project are 10 per cent samples of census records. In the case of Jamaica this provides researchers with a reasonable sample for analysis although for Saint Lucia the small size of the sample of records makes it inadequate for many analyses.

The Derek Gordon Databank and the Statistical Institute of Belize disseminate full count microdata albeit with direct identifying variables and detailed geographic information removed and recoding of other potentially disclosive variables such as occupation. Licensed release of full count microdata significantly increases the technical feasibility of re-identifying individuals in the dataset. It means that the protection of confidentiality depends much more heavily on the administrative controls, primarily the restriction of access to trusted researchers and the strong incentives for them to observe the conditions of use of the data.

3. Secure data laboratories

Secure data laboratories are used by statistical offices and by data archives in universities to provide researchers with on-site access to disclosive microdata. Data sets are made available in a secure physical and computing environment designed to prevent unauthorised access to the data and to prevent microdata or disclosive outputs from leaving the laboratory. The microdata made available in laboratories are much more lightly protected by technical disclosure control measures. Similar to the application procedure for access to licensed use files, researchers also have to apply to use data in secure data laboratories. The primary means of disclosure control is review of the analytic and research outputs produced by the researcher. Review of subsequent publications may also form a condition of access to the laboratory.

There are two Caribbean statistical offices which have created secure data laboratories: the Statistical Institute of Belize (SIB) and the General Bureau of Statistics (GBS) of Suriname. At the SIB, researchers can apply for supervised access to lightly protected census datasets containing all census records. Users have to work on a dedicated computer without internet access and any output that the researcher generates must be vetted and approved before it can leave the premises.

Any person seeking to access microdata is required to submit a completed application form, CVs for all researchers who are members of the research team, and an abstract outlining the research being undertaken. There is also a requirement that, upon completion of the research, the final report or paper is shared with the SIB. The GBS s microdata laboratory provides on-site access in a similar way although it does not provide access to full count census microdata. The datasets available from the 2004 and 2012 Suriname censuses are 10 percent samples of records.

Microdata laboratories can be expensive to maintain. The physical and computing infrastructure needs to be staffed and maintained and the staff must have the skills to carry out disclosure control checks of research outputs. Furthermore, on-site laboratories tend not to be heavily used by researchers. The requirement to pass through an application procedure combined with the requirement to carry out analysis on-site is a significant disincentive to researchers. The Derek Gordon Databank was actually set up along the lines of a data laboratory with computers for use by researchers although it actually operates more as a distributor of licensed use microdata rather than a secure data laboratory. If a secure data laboratory was to be created to serve the Caribbean, the Derek Gordon Databank at the Mona campus of the University of the West Indies in Kingston, Jamaica, is a prime location. However, whether there would be sufficient demand for such a facility to be cost effective is open to question.

4. Remote access facilities

Partly to overcome the limitations of physical laboratories, some statistical offices have developed remote access facilities. The key feature of remote access facilities is that users do not have direct access to the microdata they cannot see individual census records. However, researchers are able to submit queries which are then run on the microdata with the results being returned to the researcher. Two of the most well developed systems are the Australian Bureau of Statistics (ABS) Remote Access Data Laboratory (RADL) and Statistics Canada s Real Time Remote Access (RTRA) system. In both cases users have to be licensed to use the system. The ABS s RADL system allows users to submit programs written in the SAS, SPSS or Stata languages while Statistics Canada s RTRA system utilises SAS. Disclosure control checks are applied to the results before they are returned to the user. Sample or dummy datasets are provided to help researchers to prepare programs.

Remote access facilities overcome some of the main disadvantages of on-site laboratories. They are more cost effective to run and the user can access them from their desktop. Some investment and technical expertise is required to develop the infrastructure to offer this service. Given the capacity constraints faced by Caribbean statistical offices, the in-house development of such facilities is not likely to be a practical option in the near future. However, through collaboration with data archives, specialists in dissemination of microdata and statistical disclosure control methods, it may be possible for Caribbean statistical offices to provide access to their census data in this way. The MPC, through the IPUMS-International project, is planning to release remote access facilities which will make it possible for small countries to offer access to more detailed microdata than is currently possible. ECLAC s REDATAM software is primarily a tool for interactive tabulation and computation of basic indicators although it also provides for the online submission of queries, written in the REDATAM language, to remotely held microdata.

5. Data Archives

An important consideration for statistical offices is whether to manage the dissemination of microdata themselves, or whether to use a data archive to serve as an intermediary with the researcher. In practice, a great deal of microdata dissemination takes place through data archives. Statistical offices often find it convenient to pass at least some responsibility for microdata dissemination onto data archives. Dissemination of microdata can become very time-consuming and frustrating for statistical offices and researchers alike, particularly if procedures for dissemination are not clearly established or efficiently operated. Data archives, typically residing in universities, receive data from statistical offices and in return provide access to researchers on behalf of the statistical office through agreed procedures.

They may hold and disseminate national microdata or datasets from many countries. Such archives have played a hugely important role in broadening access to microdata, creating archives of microdata, and developing statistical disclosure control methodology.

There are several international data archives that store and disseminate microdata from surveys but relatively few that have an extensive collection of census microdata. A search of the well-known archives such as the UK Data Service based at the University of Essex, the Data Library at the University of Toronto and the Inter-University Consortium for Political and Social Research (ICPSR) did not identify any recent international census microdata. These archives generally store only national census data.

The Minnesota Population Center (MPC), through its IPUMS-International project, has the largest archive of international census microdata in the world. It works in collaboration with national statistical offices and other organizations to inventory, preserve, harmonize, and disseminate census microdata from around the world. The data are made available to qualified researchers free of charge through a web dissemination system. The IPUMS-International project provides statistical offices with a platform for worldwide dissemination of their census microdata. Data are released in the form of licensed use files. The MPC employ proven and trusted methods of disclosure control, or alternatively use the disclosure control methods of the respective national statistical office providing the data. Samples of records are most commonly 10 percent. The samples of census records are also integrated and harmonised to produce variables with consistent codes thereby facilitating cross-national and cross-temporal comparisons.

The Derek Gordon Databank of the Sir Arthur Lewis Institute of Social and Economic Studies (SALISES) at the University of the West Indies is the only data archive specialising in Caribbean social and economic microdata. It stores, documents, anonymizes (using basic techniques), and distributes data mainly to researchers in universities. Its administrative rules governing distribution and use are broadly similar to IPUMS-International, for example signed confidentiality agreements and a requirement for information about research proposals. It requires that the user of the microdata submits a copy of any published report or thesis that was produced using the data, which the databank then submits to the statistical office or other owner of the original data. Acting in liaison with the data owner, the databank also fulfils requests for special samples which it tailors to meet the needs of individual researchers, while still protecting statistical confidentiality. Of course, this is dependent on the owner depositing a very large sample or the complete census microdata. The databank has been operating since 1994 with no known breach of confidentiality.

Through collaboration with data archives such as the Derek Gordon Databank and the IPUMS-International project, statistical offices can provide greater access to their census microdata without increasing their workload. Data archives are able to specialise in microdata dissemination and methods of statistical disclosure control in a way that statistical offices cannot. Through collaboration with data archives, statistical offices will acquire greater knowledge and experience of the technical, administrative and legal issues associated with microdata dissemination and access to both advice and anonymized microdata files. This puts statistical offices in better position to handle requests for microdata where they continue to receive them, for example from researchers in other ministries of government or government commissioned researchers.

C. The legal context for release of microdata

Statistical legislation has generally been clear about the obligation to protect the confidentiality of information provided by individuals, households or other entities for the production of official statistics. With statistical offices seeking to meet the increasing demand for microdata, modern statistical legislation normally makes some provision for the release of microdata. Legislation which explicitly authorises the release of microdata ensures public confidence in the release arrangements as well as clarity, consistency and a basis for dealing with breaches of confidentiality (see for example United Nations, 2007; or Dupriez and Boyko, 2010).

Many Caribbean statistics acts have similar wording and are clear about statistical offices obligation to prevent the disclosure of any information which can be identified as belonging to an individual, undertaking or business. However, Caribbean statistical acts are outdated (Harrison, 2012) and make little mention of microdata. They in no way prohibit its release, but they do not provide a legislative framework governing release. The model statistics bill which was prepared by the CARICOM Secretariat, with support from the Inter-American Development Bank, to support countries in updating their statistics legislation does include a clause explicitly authorising the Chief Executive Officer of a statistical office to release anonymized microdata for the purpose of research.

Statistical offices have tended to be extremely conservative in assessing the risks of disclosure and to err on the side of not releasing the data. Faced with the difficulty of making an accurate assessment of the real risk of a breach of confidentiality, the simplest and easiest way to avoid any disclosure risk is not to release any microdata. Yet in reality, instances of disclosure arising from the release of census microdata are almost unheard of.

One of the values of authorising legislation is that the balance, or trade-off, between protecting confidentiality and facilitating access to microdata is recognised in legislation. This encourages statistical offices to approach the issue in a more balanced way rather than adopting an extremely conservative approach or a presumption against the release of microdata should there be even the slightest theoretical possibility of disclosure. New legislation would certainly help to encourage and facilitate release of microdata to researchers in the Caribbean. However, steps to expand access to microdata need not wait for new legislation. The release of anonymized microdata for research purposes is a very well established international practice and the value and legitimacy of microdata release is backed by a recent policy statement of Caribbean Directors of Statistics (the SCCS in 2010). Existing laws do not prohibit the release of microdata provided the careful application of disclosure control procedures ensure that information belonging to individuals, undertakings or businesses cannot be identified.

II. The creation of microdata release files for Grenada and Trinidad and Tobago

This chapter outlines an approach to the anonymization of census microdata which shares many similarities with that used by some statistical offices and the IPUMS-International project (see for example: US Census Bureau, 2015; ONS, 2012; McCaa, Ruggles, Sobek, 2010). This constitutes a relatively practical and pragmatic application of disclosure control methods which is widely applicable to Caribbean census datasets. It has been applied here to the most recent census datasets from Grenada and Trinidad and Tobago.

A. Caribbean census datasets

The 2010 round of Population and Housing Censuses in the Caribbean was coordinated through the Regional Census Co-ordinating Committee (RCCC) organised by the CARICOM Secretariat. One of the areas in which this coordination is most evident is in the questionnaires of the respective countries and territories. Census questionnaires were shaped by common core questions, concepts and definitions.

There were core questions in the following areas for individuals: personal characteristics; migration (birthplace and residence); disability; education; training; economic activity (15 years and over); marital and union status (15 years and over); fertility (females 14+); access to the internet; income; and location on census night. Core questions on housing included: characteristics of occupied buildings; characteristics of occupied dwelling units and land tenancy; facilities available for use; international migration; the environment and crime. Non-core questions included further elaboration of the areas listed and health related issues. While there is a substantial degree of harmonisation, national variations are still evident and the questionnaires are not strictly harmonised in respect of every question or detail.

B. The creation of microdata release files

1. Removal of direct identifying variables

A first, and relatively straightforward, step in the anonymization of a census dataset is the removal of direct identifying variables such as names, addresses and telephone numbers. In addition to direct identifying variables, census datasets often contain variables which relate to administrative aspects of the census taking process. Examples of such variables include barcodes and questionnaire numbers which clearly have no relevance to researchers and should be removed. Some variables providing information about the data collection and the processing of the data could however be useful researchers, for example flags to indicate where values have been estimated or imputed. These variables should be considered for retention (taking into account any impact on disclosure risk).

2. Sampling of records

Sampling of records is one of the primary ways in which the risk of re-identification of individuals or households is reduced. Providing researchers with a sample rather than full census datasets reduces disclosure risk simply by reducing the number of records which are made available to potential intruders. Many of the measures of disclosure risk used in this analysis are related in a proportional way to the sampling fraction which has been used to create the sample of records: the lower the sampling fraction, the lower the disclosure risk. Sampling also reduces the incentive for an intruder to attempt to re-identify individuals because they know *a priori* that there is only a small probability that records belonging to particular individuals or households are present in the microdata release file.

Sampling complicates the task of a nosy neighbour because even if they are able to find a record in the sample which strongly resembles a known individual or household, and even if that record is unique in the sample, this is not sufficient to establish that the record does in fact belong to that individual or household. To verify this, it is necessary to rule out the possibility that there are other members of the population who also have those same characteristics, something that the nosy neighbour has no obvious means to achieve.

Datasets which contain information about households, together with the information about the individuals within those households, are the most useful and flexible for researchers because they permit analysis of relationships between individual and household characteristics. With this in mind, it is households which are sampled, with information about all individuals in sampled households being included in the release file.

The most straightforward way to obtain a random sample of households is to use systematic random sampling with equal probability, for example selecting every tenth household (from a random start) to obtain a 10 per cent sample of households (see United States Census Bureau (2013) and McCaa and others (2006)). The dataset should be sorted prior to the selection of the sample as the sorting variables will provide implicit stratification of the sample. This means that estimates derived from the sample will have much greater precision than would be achieved from a simple random sample of households.

There is a choice to be made concerning the variables for sorting. These variables should define strata within which there is a high degree of homogeneity with respect to characteristics of major interest. Here, the geographic variables have been used to provide a detailed stratification with datasets sorted, for example, by municipality, community and enumeration district. The rationale for this is that many variables of interest to researchers will be related to geography and systematic random samples, by ensuring a representative geographic distribution of sampled cases, are equivalent to extremely fine geographic stratification with proportional weighting (McCaa and others, 2006). Samples created in this way are exceptionally representative of the territory of a country and therefore should be representative of most social and demographic variables.

A further advantage of systematic random sampling is that the weighting of sample records to produce population estimates is also simple. In the case of a 10 per cent sample, all persons and households would be assigned a weight of 10 (where census records have unit non-response weights,

these sample weights are combined with the unit non-response weights). It is necessary to provide some additional protection to large households against the risk re-identification. A simple way of doing this is simply to re-sample any households above a certain size threshold although age perturbation could also be used to protect the identity of these households (US Census Bureau, 2013).

3. Analysis of disclosure risk, the sampling fraction, and recoding of indirect identifying variables

A file of census microdata records for release to researchers is created based on an analysis of disclosure risk which informs the design of the file. Decisions about the sampling fraction, the level of geographic information included in the file, and the coding of other indirect identifying variables should be based on an analysis of the implications of each of these decisions for the risk of disclosure. Although direct identifying variables are dropped from the dataset, an intruder could still attempt to re-identify individuals via the indirect identifying variables. This approach to analysing disclosure risk assesses how successful such attempts are likely to be.

Central to the analysis of disclosure risk is the concept of a key, that is, the variables that an intruder might use to re-identify individuals. In this scenario, the intruder is presumed to have information about the population in an external database containing some variables which are the same, or similar, to some of the variables contained in the census microdata release file. The intruder is presumed to match the census microdata release file with the external database in an attempt to establish the identity of the persons or households corresponding to the individual census records (note that this would be a breach of the terms of most data access agreements). With a unique match, the intruder can establish the identity of the person or household corresponding to a census record. The intruder then has access to that person s full census record in the microdata file and disclosure is considered to have taken place. In practice, there is uncertainty about the information that an intruder possesses and therefore this analysis presents results for a range of different sets of key variables representing different assumptions about the data that the intruder will have access to. These assumptions range from the very weak, assuming the intruder only has data on the age and sex of the population, to the very strong/conservative assumption that the intruder has information for 10 separate social/demographic characteristics.

This approach to analysing disclosure risk needs to be understood as providing a very conservative assessment of risk for several reasons. It is unlikely that the information in an external database is classified in exactly the same way as the information in the census. It is equally unlikely that the information refers to the same point in time as the census. Furthermore, this analysis presumes that all data items in both the census and the external database are recorded without measurement error. Also, by treating the complete set of census records as if it was the population, and ignoring census non-response, the analysis further over-estimates the risk of disclosure. These measures of risk therefore, in reality, are better understood as upper bounds for the level of disclosure risk.

Multiple samples of records were produced using different sampling fractions, inclusion or exclusion of first level geographic codes, and different methods of recoding the identifying variables (or top coding in the case of age). The analysis provides measures of the impact on disclosure risk of each decision which is taken in the design of a microdata release file. This provides a firm basis for assessing the *relative* impact of these decisions on disclosure risk. However, because of the limitations of this analysis, deciding what is actually an acceptable level of risk for a microdata file to be released to a given set of users still comes down to judgement. This judgement about risk will differ according to whether microdata is being released in the form of a public use file, a licensed use file, made available through a secure laboratory or remote access facilities.

The indicators of disclosure risk are calculated based on measures of the rareness or uniqueness of cases in either the full set of census records, or the samples of records which would constitute a microdata release file. While the different measures of risk provide slightly different perspectives, or are based on slightly different assumptions, in general they tend to respond in predictable ways to changes in the design of the sample of records. For example, recoding a variable such as religion by combining categories will reduce disclosure risk on all the measures.

Most focus has been placed on the number (and proportion) of population uniques which are present in the sample of records because these are the cases at greatest risk of re-identification. They are the records that an intruder with access to some external database could possibly match uniquely.

The results for a sample of records from the Grenada 2011 census (a 10 per cent sample including geographic codes (parish) and single year of age up to 75+) are shown in Table 1. An intruder armed with information about sex, age and parish of residence for the population could not realistically hope to re-identify any census records. In this 10 per cent sample of records, not one record could be uniquely re-identified using these three variables. If the intruder additionally had information about marital status and occupation (2 digit) they could still only uniquely match 490 records, 5 per cent of the sample (less than 0.5 per cent of the population). If an intruder were to additionally have information on ethnicity and citizenship, a key of seven variables, this proportion increases to 16 per cent of the sample (1,542 persons) or 1.5 per cent of the population. If the intruder additionally has industry (2 digit) the disclosure risk increases further. It is difficult to conceive of an intruder being able to match systematically using more than eight variables but nevertheless were an intruder to have information about a person s level of education and place of birth, they could uniquely re-identify 46 per cent of the sample (or 4,548 persons). Again it is worth emphasising that this analysis is based on conservative assumptions about what information might be available to the intruder, and so these statistics indicate the maximum number of records that could possibly be re-identified by an intruder in a worst case scenario.

For each of the columns in Table 1 showing the number of sample and population uniques, the adjacent column shows the number of these uniques which are confounded by other cases with missing values on the key variables. These are records which are probably unique (and are counted as unique) but due to missing data for other sample records, there is some uncertainty about whether they are in fact unique. This is relevant because it illustrates how missing data, by introducing uncertainty, provides some level of protection against re-identification. For short and mid-length keys, there are a significant proportion of uniques about which there exists this uncertainty (see also the technical notes which appear in the Annex).

Table 2 compares measures of disclosure risk for this and three alternative designs of microdata release file. Using the number of population unique cases in the sample as a single indicator of risk, Figure 1 compares these four possible designs. Based on a key of eight variables, in the first sample of records (a 10 per cent sample including parish and single year of age up to 75+) there are 2,697 population unique records in the sample, 27.5 per cent of the sample or around 2.5 per cent of the population.

A sample of 10 per cent of census records from the Grenada dataset yields a sample of 9,825 records which is very small. A 20 per cent sample of records would be more useful to analysts (19,240 records) however the disclosure risk for the 10 per cent sample of records is already quite high. Figure 1 shows that the sampling fraction could be increased to 20 per cent, without any increase in disclosure risk, if the parish variable was dropped from the data (at least based on a key of eight variables).

The 20 per cent sample without the parish variable actually has a lower disclosure risk for shorter keys but a higher disclosure risk for longer keys (because the addition of more variables to the key eventually outweighs the effect of having removed the parish variable). This is an illustration of how geography is a very powerful identifying variable and significant reductions in disclosure risk are obtained by reducing (or in this case eliminating) geographic detail (with obvious costs to researchers).

If this level of disclosure risk is deemed too high, two alternatives which reduce the level of disclosure risk would be a 10 per cent sample which excludes the parish variable or a 20 per cent sample which not only excludes the parish variable but also groups age into five year age bands. Each of these samples would reduce the disclosure risk, or the number of risky records, by about half.

Table 1
Measures of disclosure risk for a 10 percent sample of records from the Grenada 2011 Census (9 825 persons)

Number of key variables	Key Variables	Number of sample uniques (SU)	SU confounded by other sample records with missing values[a]	Percentage of SU in sample	Expected percentage of correct matches for SU[b]	Percentage of SU that are also PU	Population uniques (PU) in sample[c]	PU confounded by other sample records with missing values[d]	Percentage of PU in sample	Sample records not k-anonymous (k=3) in the population[e]	Global risk[f]
2	Age, Sex	0	0	0.0	0.0	0.0	0	0	0.0	0	15
3	Age, Sex, Parish	102	0	1.0	6.3	0.0	0	0	0.0	0	124
4	Age, Sex, Parish, Marital Status	673	89	6.9	26.3	10.4	70	10	0.7	141	363
5	Age, Sex, Parish, Marital Status, Occupation (2 digit)	2 522	551	25.7	37.6	19.4	490	92	5.0	928	1 258
6	Age, Sex, Parish, Marital Status, Occupation (2 digit), Ethnicity	3 317	508	33.8	46.2	29.1	966	97	9.8	1 545	1 867
7	Age, Sex, Parish, Marital Status, Occupation (2 digit), Ethnicity, Citizenship	4 133	519	42.1	53.4	37.3	1 542	118	15.7	2 291	2 550
8	Age, Sex, Parish, Marital Status, Occupation (2 digit), Ethnicity, Citizenship, Industry (2 digit)	4 931	131	50.2	68.2	54.7	2 697	35	27.5	3 566	3 642
9	Age, Sex, Parish, Marital Status, Occupation (2 digit), Ethnicity, Citizenship, Industry (2 digit), Place of birth	5 653	185	57.5	74.5	63.4	3 581	51	36.5	4 436	4 470
10	Age, Sex, Parish, Marital Status, Occupation (2 digit), Ethnicity, Citizenship, Industry (2 digit), Place of birth, Education	6 580	151	67.0	78.5	69.1	4 548	43	46.3	5 397	5 412

Source: Economic Commission for Latin America and the Caribbean (ECLAC), on the basis of data provided by the Central Statistical Office of Grenada.

[a] These are individuals that appear to be sample uniques (and are counted as sample uniques) but might not be sample unique depending on the unknown missing values in other cases on the key variables.
[b] The expected proportion of correct matches among sample uniques if the sample of records was matched against population data, using the key variables, with one-to-many matches assigned to a population record at random.
[c] These sample records are unique in the population on the key variables (if the sample of records was matched against population data, using the key variables, these records could be matched uniquely).
[d] These are individuals that appear to be population uniques (and are counted as population uniques) but might not be population unique depending on the unknown missing values in other cases on the key variables.
[e] Records which are not k-anonymous (k=3) in the population are those which are either unique on the key variables or share the same key as no more than one other record.
[f] The global risk is the sum for the sample of all the individual record risks. It can be understood as an expected number of correct matches if all sample records are matched with population data, using the key variables, with one-to-many matches assigned to a population record at random.

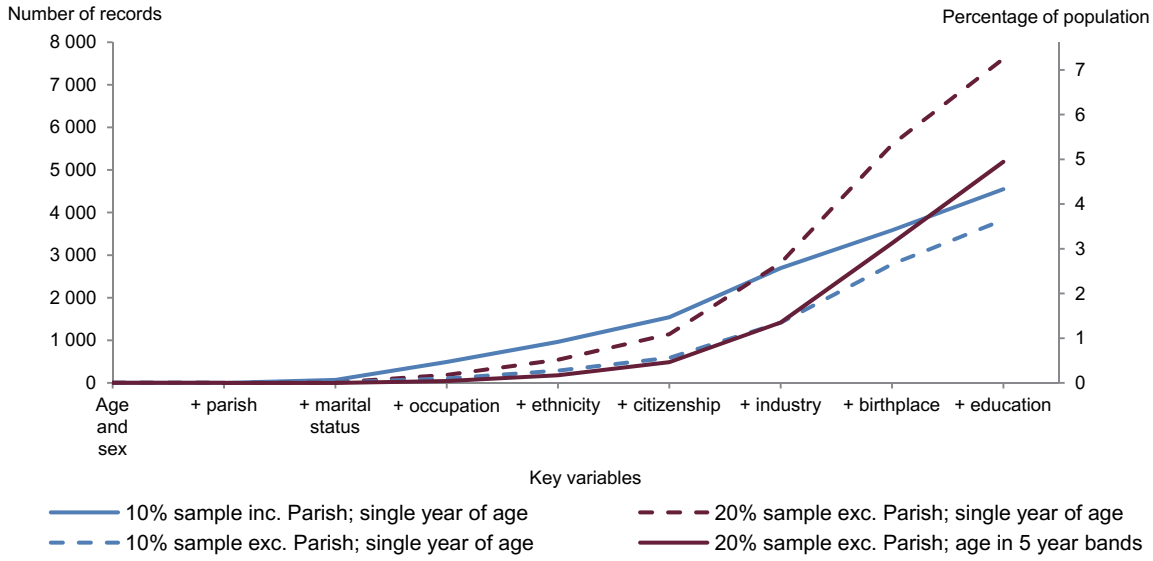
Table 2
Measures of disclosure risk for samples of anonymized records from the Grenada 2011 Census

Key variables[a]	Number of sample uniques (SU)	SU confounded by other sample records with missing values	Percentage of SU in sample	Expected percentage of correct matches for SU	Percentage of SU that are also PU	Population uniques (PU) in sample	PU confounded by other sample records with missing values	Percentage of PU in sample	Sample records not k-anonymous (k=3) in the population	Global risk
10 percent sample of records including the parish variable; single year of age (9 825 person records)										
Age and sex	0	0	0.0	0.0	0.0	0	0	0.0	0	15
+ parish	102	0	1.0	6.3	0.0	0	0	0.0	0	124
+ mar. status	673	89	6.9	26.3	10.4	70	10	0.7	141	363
+ occupation	2 522	551	25.7	37.6	19.4	490	92	5.0	928	1 258
+ ethnicity	3 317	508	33.8	46.2	29.1	966	97	9.8	1 545	1 867
+ citizenship	4 133	519	42.1	53.4	37.3	1 542	118	15.7	2 291	2 550
+ industry	4 931	131	50.2	68.2	54.7	2 697	35	27.5	3 566	3 642
+ birthplace	5 653	185	57.5	74.5	63.4	3 581	51	36.5	4 436	4 470
+ education	6 580	151	67.0	78.5	69.1	4 548	43	46.3	5 397	5 412
10 percent sample of records excluding the parish variable; single year of age (9 825 person records)										
Age and sex	0	0	0.0	0.0	0.0	0	0	0.0	0	15
~~+ parish~~	0	0	0.0	0.0	0.0	0	0	0.0	0	15
+ mar. status	94	73	1.0	22.3	9.6	9	8	0.1	16	66
+ occupation	696	490	7.1	29.8	15.1	105	65	1.1	188	350
+ ethnicity	1 254	551	12.8	38.8	23.0	288	83	2.9	476	676
+ citizenship	1 965	684	20.0	45.3	30.0	589	138	6.0	888	1 108
+ industry	3 293	177	33.5	59.0	43.0	1 416	47	14.4	2 020	2 151
+ birthplace	4 997	357	50.9	68.7	55.7	2 785	45	28.4	3 642	3 713
+ education	6 077	364	61.9	73.9	63.0	3 828	58	39.0	4 717	4 757
20 percent sample of records excluding the parish variable; single year of age (19 240 person records)										
Age and sex	0	0	0.0	0.0	0.0	0	0	0.0	0	30
~~+ parish~~	0	0	0.0	0.0	0.0	0	0	0.0	0	30
+ mar. status	70	67	0.4	33.6	15.7	11	11	0.1	22	119
+ occupation	745	608	3.9	42.4	24.8	185	142	1.0	339	664
+ ethnicity	1 589	889	8.3	51.1	34.2	544	228	2.8	905	1 307
+ citizenship	2 719	1 199	14.1	57.3	42.1	1 144	375	6.0	1 754	2 171
+ industry	5 211	388	27.1	68.4	54.0	2 816	157	14.6	3 983	4 238
+ birthplace	8 535	641	44.4	76.9	65.5	5 591	181	29.1	7 226	7 358
+ education	10 634	699	55.3	81.1	71.6	7 611	207	39.6	9 308	9 382
20 percent sample of records excluding the parish variable with age banded in 5 year groups (19 240 person records)										
Age and sex	0	0	0.0	0.0	0.0	0	0	0.0	0	8
~~+ parish~~	0	0	0.0	0.0	0.0	0	0	0.0	0	8
+ mar. status	12	11	0.1	23.1	0.0	0	0	0.0	2	33
+ occupation	219	194	1.1	39.9	20.1	44	38	0.2	106	225
+ ethnicity	624	452	3.2	47.3	28.9	180	113	0.9	331	530
+ citizenship	1 274	805	6.6	55.1	38.3	488	276	2.5	795	1 035
+ industry	2 930	369	15.2	63.3	48.5	1 420	124	7.4	2 048	2 374
+ birthplace	5 796	598	30.1	70.4	56.5	3 277	192	17.0	4 566	4 790
+ education	7 990	714	41.5	76.3	65.0	5 192	262	27.0	6 669	6 818

Source: Economic Commission for Latin America and the Caribbean (ECLAC), on the basis of data provided by the Central Statistical Office of Grenada.

[a] Measures of disclosure risk are calculated for a range of different sets of key variables which reflect increasingly conservative assumptions about the information that an intruder might have about the population. The weakest key consists of just two variables, age and sex, while the strongest consists of ten separate pieces of information about individuals.

Figure 1
Number of population uniques in samples of records from the 2011 Census of Grenada with respect to selected sets of key variables[a]



Source: Economic Commission for Latin America and the Caribbean (ECLAC), on the basis of data provided by the Central Statistical Office of Grenada.

[a] Key variables are added sequentially so the first key is simply age and sex; the second key is age, sex and municipality; the third key is age, sex, municipality and marital status and so on.

Table 3 shows a similar analysis of four different microdata release files for the Trinidad and Tobago 2011 Census. Figure 2 shows the number of population uniques in those samples. Take, for example, a key consisting of eight variables (age, sex, municipality, marital status, religion, ethnicity, level of education and activity status). In a 10 per cent sample of records including the municipality variable and single year of age (up to 90+), approximately 36,000 out of 116,000 records are unique in the population on that key (more precisely they are unique among households responding to the census). This is equivalent to around 30 per cent of the sample or 3 per cent of the population. If it was felt that there were too many high risk, population unique records in this sample, various steps can be taken to reduce the number of such records, and Figure 2 shows three examples of slightly different sample designs. Reducing the sampling fraction by half (to 5 per cent), not surprisingly, reduces the number of population unique cases in the sample by half. This reduces the disclosure risk by half. Recoding age into five year bands, rather than single year of age, has a roughly similar effect. The adjustment which has the largest effect is to remove the municipality variable entirely from the dataset. This reduces the number of population unique records in the sample by about two-thirds.

This last step in particular is not desirable. In a country of over 1.3 million persons it should be possible to provide at least some geographic information in a sample of census records. This illustrates the point that was made above, that there are limits to the application of methods of data reduction. If taken too far they ultimately destroy the utility of the dataset.

This analysis could equally have been carried out using some of the other measures of disclosure risk, for example the number of records in the sample failing the k-anonymity (k=3) threshold or the global risk. While these measures are on a different scale they generally lead to the same conclusions about the relative impacts of sampling fractions and coding decisions on disclosure risk.

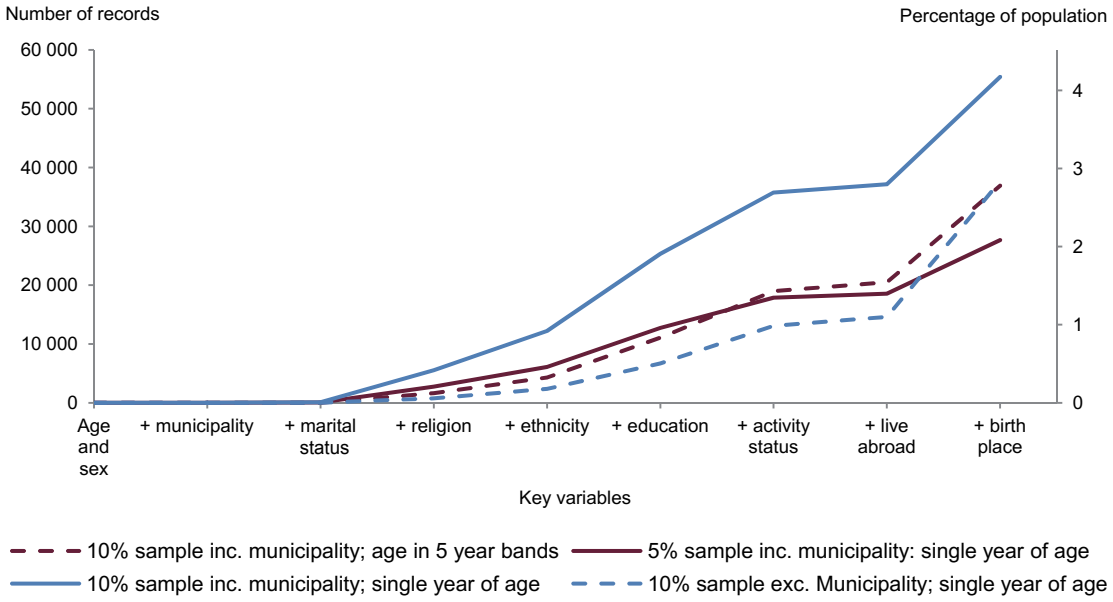
Table 3
Measures of disclosure risk for samples of anonymized records from the Trinidad and Tobago 2011 Census

Key variables[a]	Number of sample uniques (SU)	SU Confounded by other sample records with missing values	Percentage of SU in sample	Expected percentage of correct matches for SU	Percentage of SU that are also PU	Population uniques (PU) in sample	PU confounded by other sample records with missing values	Percentage of PU in sample	Sample records not k-anonymous (k=3) in the population	Global risk
10 percent sample of records including municipality; single year of age (116 340 person records)										
Age and sex	0	0	0.0	0.0	0.0	0	0	0.0	0	18
+municipality	41	0	0.0	5.9	0.0	0	0	0.0	0	274
+ mar. status	1 655	1 120	1.4	19.3	5.6	92	56	0.1	211	1 047
+ religion	26 631	2 223	22.9	38.2	20.7	5 514	230	4.7	9 543	13 802
+ ethnicity	43 329	3 949	37.2	45.5	28.2	12 223	631	10.5	19 675	24 091
+ education	63 390	4 126	54.5	56.3	40.0	25 331	850	21.8	37 115	40 177
+ activity st.	72 373	4 977	62.2	63.6	49.4	35 753	1 463	30.7	48 570	50 118
+ live abroad	73 209	4 840	62.9	64.7	50.8	37 154	1 430	31.9	49 879	51 323
+ birthplace	86 259	5 468	74.1	74.7	64.3	55 419	1 957	47.6	68 030	67 624
10 percent sample of records excluding municipality; single year of age (116 340 person records)										
Age and sex	0	0	0.0	0.0	0.0	0	0	0.0	0	18
~~+municipality~~	0	0	0.0	0.0	0.0	0	0	0.0	0	18
+ mar. status	45	45	0.0	14.6	0.0	0	0	0.0	3	79
+ religion	4 956	920	4.3	31.7	15.2	752	61	0.7	1 357	2 703
+ ethnicity	11 469	2 014	9.9	37.1	20.7	2 377	229	2.0	3 942	6 156
+ education	24 774	4 078	21.3	43.9	27.1	6 707	560	5.8	10 741	13 785
+ activity st.	35 807	6 701	30.8	52.0	36.6	13 094	1 522	11.3	19 215	21 761
+ live abroad	37 264	6 560	32.0	54.1	39.2	14 614	1 519	12.6	20 848	23 278
+ birthplace	70 942	17 634	61.0	65.0	53.1	37 658	4 611	32.4	49 241	49 687
10 percent sample of records including municipality; age in 5 year bands (116 340 person records)										
Age and sex	0	0	0.0	0.0	0.0	0	0	0.0	0	4
+municipality	10	0	0.0	6.5	0.0	0	0	0.0	0	59
+ mar. status	242	198	0.2	16.2	3.7	9	8	0.0	27	247
+ religion	9 368	1 332	8.1	34.4	17.5	1 641	105	1.4	2 942	5 068
+ ethnicity	18 746	2 631	16.1	40.0	23.0	4 310	319	3.7	7 292	10 294
+ education	35 478	4 380	30.5	47.9	31.2	11 083	699	9.5	17 369	20 764
+ activity st.	46 419	6 383	39.9	55.9	40.9	18 996	1 586	16.3	27 262	29 789
+ live abroad	47 660	6 232	41.0	57.5	43.0	20 495	1 556	17.6	28 814	31 227
+ birthplace	64 927	9 368	55.8	68.1	56.9	36 909	3 087	31.7	47 493	47 952
5 percent sample of records including municipality; single year of age (58 189 person records)										
Age and sex	0	0	0.0	0.0	0.0	0	0	0.0	0	9
+municipality	87	0	0.2	3.8	0.0	0	0	0.0	0	137
+ mar. status	1 684	882	2.9	12.2	2.6	44	17	0.1	104	518
+ religion	19 820	1 290	34.1	28.6	13.8	2 742	63	4.7	4 776	6 902
+ ethnicity	29 080	1 992	50.0	36.8	20.9	6 090	196	10.5	9 905	12 082
+ education	38 438	1 760	66.1	49.0	33.1	12 731	239	21.9	18 616	20 219
+ activity st.	41 959	1 945	72.1	57.1	42.6	17 886	429	30.7	24 379	25 193
+ live abroad	42 267	1 888	72.6	58.1	43.9	18 564	423	31.9	25 031	25 777
+ birthplace	47 382	1 977	81.4	69.6	58.4	27 656	553	47.5	34 048	33 944

Source: Economic Commission for Latin America and the Caribbean (ECLAC), on the basis of data provided by the Central Statistical Office of Trinidad and Tobago.

[a] Measures of disclosure risk are calculated for a range of different sets of key variables which reflect increasingly conservative assumptions about the information that an intruder might have about the population. The weakest key consists of just two variables, age and sex, while the strongest consists of ten separate pieces of information about people.

Figure 2
Number of population uniques in samples of records from the 2011 Census of Trinidad and Tobago with respect to selected sets of key variables[a]



Source: Economic Commission for Latin America and the Caribbean (ECLAC), on the basis of data provided by the Central Statistical Office of Grenada.

[a] Key variables are added sequentially so the first key is simply age and sex; the second key is age, sex and municipality; the third key is age, sex, municipality and marital status and so on.

4. Data swapping

It is clear from the preceding discussion that however the key variables are recoded and whatever sampling fraction is chosen there will still be some risky records remaining in the microdata release file which are potentially identifiable via key variables. No amount of global recoding can avoid this without rendering the file useless to most researchers. At this stage it is necessary to resort to perturbative methods of disclosure control. The intention here is to introduce a small level of random perturbation of the dataset that is not so significant as to affect the results of any analysis that might be carried out, but nevertheless introduces sufficient uncertainty about whether values in the dataset are in fact the real values or have been perturbed. In this way a potential intruder is not able to claim with any certainty that the identity of a person or household has been established.

The preceding analysis provides a good indication of which records are at greatest risk of re-identification. It therefore makes sense that any perturbation of records should be directed towards the records with the highest risk of re-identification

Data swapping has been used by national statistical offices to protect confidentiality of both tabulated census data and census microdata (United States Census Bureau, 2015; Ito and Hoshino, 2014; Office for National Statistics, n/d; McCaa, Ruggles and Sobek, 2010). When data swapping is applied to census data, normally, only geographic information is swapped. The place of residence of individuals and households is likely to be involved in any attempt at re-identification and so introducing uncertainty into geographic information makes the task of an intruder harder and specifically, makes it possible to refute any claim that an individual or household in the microdata release file has been re-identified with certainty. For most analyses, swapping the location of a small proportion of whole households has a relatively minor impact on the dataset. Note that this method is only relevant where a geographic variable is included in the microdata release file.

It involves selecting a set of households for swapping, and for each of those households identifying a matching household in another geographic area. The households should match at least in respect of household size, numbers of male and female children, number of male and female adults, and numbers of male and female older persons. Having identified matching pairs of households, the geographic codes are swapped between them. In countries which have applied this method (larger countries than those in the Caribbean), matches are identified at relatively close geographic proximity, for example within the same county, so at that level (and more aggregated geographies) there is no effect. In the Caribbean, due to the small size of the countries and territories, census datasets made available to researchers are likely to include either one level of geography or no geographic codes at all. In the data swapping carried out as part of this analysis, the swapping algorithm searched for matches in any part of the country (discarding possible matches from the same geographic area).

In order to provide greater protection for records at higher risk of re-identification, those households containing higher risk individuals are more likely to be selected for swapping. Households are selected for swapping with probability proportional to the household s risk of re-identification (see Annex). In practice this means that many of the records which are swapped are population uniques.

With this targeting of the record swapping, it is clear that swapping a relatively small proportion of the households in the sample of records can still represent a significant proportion of the most risky records. The swapping therefore introduces significant uncertainty about the true location (and therefore identity) of precisely the records most at risk of re-identification. The percentage of households which are swapped is generally not made public. Uncertainty about the exact nature of the disclosure control methods which have been used to create a microdata release file is generally regarded as an additional source of protection against re-identification.

5. Recoding of non-identifying variables

Non-identifying variables are those which cover topics such as health, disability, fertility etc. and most of the housing variables. Individuals and households can still be rare or unique in the population with respect to these variables, and therefore they also create a risk of re-identification for some records. These variables need to be checked and where necessary recoded so that there are an acceptable minimum number of cases within each category. The IPUMS confidentiality protocols commonly apply a minimum threshold of 250, meaning that for any social characteristic represented in the form of a categorical variable, the minimum number of individuals for any category of that variable is 250. Where the number of individuals falls below 250 that characteristic is re-coded as missing, suppressed or aggregated (McCaa and others, 2014). The 250 persons (or 60 households) threshold has also been used here.

There is some scope for relaxing the thresholds of 250 persons or 60 households particularly for some variables or categories of variables. For example, both censuses asked how many hours respondents worked in the previous week. It is difficult to believe that a person could be identified in a microdata release file, published two or three years after the census, based on information about how many hours they had worked in the week before the census. Both censuses asked about people s reasons for doing certain things, for example reason for not seeking work or reason for returning to live in Grenada . These questions are asking people to give, effectively, an opinion. In the same way, it is very difficult to see how people could be identified from their responses to these questions.

It is clear that in the Grenada census more variables have to be removed or recoded and more categories suppressed than in the Trinidad and Tobago census (Tables 4, 5 and 6). Grenada is 12-13 times smaller than Trinidad and Tobago and so application of the 250 persons/60 household threshold necessarily leads to greater suppression in this dataset. Even so, 250 persons still represents less than 0.25 per cent of the population of Grenada. This suppression is important to avoid persons with rare population characteristics from being identified in the microdata. It probably has relatively little impact on data utility: once the sample of records is taken, inferences about characteristics which occur in less than 0.25 per cent of the population would not be reliable anyway.

Table 4
Variables to be removed from the microdata files

Country	Variables removed
Trinidad and Tobago	Community, Enumeration district, Building number, Dwelling unit number, Household number (household variables); and date of birth
Grenada	Phone number, Parish, Village, Enumeration District, House Number, Variables related to some crimes (Murder, Kidnapping, Shooting and Rape), Variables relating to deaths in households where number of deaths > 1 (household variables); and person identification number, date of birth, usual address, place of residence five years ago, place of residence ten years ago, some variables related to aids (crutches, brailler, adapted car, prosthesis, orthopaedic shoes, hearing aid, other), some variables related to health conditions (carpal tunnel syndrome, lupus, HIV), name of school attended, place of work, date of last live birth, variables relating to infant deaths, location on census night (person variables)

Source: Economic Commission for Latin America and the Caribbean (ECLAC), on the basis of data provided by the Central Statistical Office of Trinidad and Tobago and the Central Statistical Office of Grenada.

Note: The variables removed are either direct identifying variables, geographic identifiers, or variables which have been suppressed because they measure rare characteristics and recoding is not possible.

Table 5
Recoding of variables for anonymization, Grenada 2011 Census

Variables	Recodes, top codes and bottom codes
Household variables	
Type of Dwelling	Recode Barracks, Outhouse, Group Dwelling, Improvised Housing Unit → other
Housing tenure	Recode Rent from Government → other
Rent	Group
Mortgage	Group
Wall material	Recode Stone, Brick → other
Roof	Recode Shingle (other) → other
Water supply	Recode Truck borne → other
Source of lighting	Recode Solar → other
Fuel for cooking	Recode Solar → other
Number of rooms	Top code 12+
Number of bedrooms	Top code 8+
Garbage disposal	Recode Dumping, Burying → other
Desktop computers	Top code 3+
Laptop computers	Top code 4+
Vehicles	Top code 5+
Age of persons dying in last 12 months	Five year age groups up to 85+
Number of migrants	Top code 4+
Person variables	
Relationship to head of household	Recode Spouse/partner of child of head/spouse/partner → other relative; recode domestic employee → other non-relative.
Age	Single year of age up to 75+ or 5 year age bands up to 90+
Ethnicity	Recode Indigenous people, Chinese, Portuguese, Syrian/Lebanese, Hispanic → other
Religion	Recode Baha'i, Hindu, Moravian, Salvation Army, Lutheran → other
Place of birth	Recode to parish; all countries except United States of America, United Kingdom, Canada, Trinidad and Tobago, Guyana → other
Year came/returned to Grenada/Parish	Group years
Place of previous residence	Recode to parish; all countries except United States of America, United Kingdom, Canada, Barbados, Trinidad and Tobago → other
Reason for returning to Grenada	Recode Involuntary return, Homesick → other
First citizenship, Second citizenship	Recode all countries except United States of America, United Kingdom, Canada, Trinidad and Tobago, Grenada and Guyana → other
Disabilities	Merge the categories 'Yes, lots of difficulty' with 'cannot do at all'
Health insurance plan	Recode Endowment with health and School accident insurance → other
Type of school	Recode Daycare, Special education, Post Primary, Home schooling, Adult education → other

Table 5 (concluded)

Highest level of education	Merge the categories 'Tertiary level – Masters' and 'Doctorate level programmes'
Highest examination	Recode Cambridge School Certificate, Post Graduate Certificate, Post Graduate Diploma, Higher Degree → other
Duration of training	Group durations
Type of certification	Recode Advance Diploma, Associated degree, Post Graduate degree → other
Category of work	Recode Apprentice/Learners, Unpaid worker, Unpaid family worker → other
Livelihood	Merge the categories 'Disability benefits' with 'Social security'
Age at first union	Group ages above 40
Total M/F children born/still alive	Top code 7+
Age at first birth/last birth	Bottom code <=15, group ages over 35
Year of last live birth	Group years before 1980
Live births (M/F) in last 12 months	Recode >1 as 'not stated'

Source: Economic Commission for Latin America and the Caribbean (ECLAC), on the basis of data provided by the Central Statistical Office of Grenada.

Table 6
Recoding of variables for anonymization, Trinidad and Tobago 2011 Census

Recoded variables	Recodes, top codes and bottom codes
Household variables	
Wall material	Recode Thatch/ Makeshift → other
Rent	Group
Lighting	Recode Gas → other
Cooking fuel	Recode Solar energy → other
Number of migrants from HH	Top code 5+
Person variables	
Age	Single year of year up to 90+ or five year age groups up to 95+
Religion	Recode to Anglican, Baptist, Pentecostal, Roman Catholic, Seventh Day Adventist, Brethren, Hinduism, Islam, Presbyterian, Rastafarian, Other religion, None
Place of Birth	Recode to municipality or Barbados, Canada, China, Dominica, Germany, Grenada, Guyana, India, Jamaica, Nigeria, Philippines, Saint Lucia, Saint Vincent and the Grenadines, United Kingdom, United States of America, Venezuela (Bolivarian Republic of) and other
Duration of residence	Bottom code <= 1955
Usual residence	Recode to municipality
Residence in 2000	Recode to municipality
Country of last residence	Recode to Barbados, Canada, Grenada, Guyana, Jamaica, United Kingdom, United States of America, Venezuela (Bolivarian Republic of) and other
Year left Trinidad and Tobago	Merge the categories '2010' and '2011'
Type of school	Recode Home school → other
Employment status	Other unpaid worker/employee → other
Children/males/females born/surviving	Top code 8+
Age at first/last birth	Bottom code <=15, Top code 40+
Live births (M/F) in last 12 months	Recode >1 as 'not stated'
Infant (M/F) deaths in last 12 months	Recode >1 as 'not stated'
Still births in last 12 months	Recode >1 as 'not stated'

Source: Economic Commission for Latin America and the Caribbean (ECLAC), on the basis of data provided by the Central Statistical Office of Trinidad and Tobago.

C. Release conditions: licensed or public use

The steps presented above can be used to anonymize Caribbean census datasets and the approach used here is similar, at least in broad terms, to the approach used in the anonymization of census data by the United States Census Bureau, the United Kingdom s Office for National Statistics and the Minnesota Population Center s IPUMS International series. It is appropriate for the creation of either public use files or licensed use files.

In the case of Grenada, under licensed conditions, a 20 per cent sample of records could reasonably be released (excluding the parish variable). This dataset would only be distributed to licensed researchers who would have strong disincentives to misuse the data. It should be emphasised that even if a researcher did try to re-identify individuals, they would face considerable difficulties and challenges: finding sufficient information with which to attempt a matching exercise; different concepts, classifications, reference dates; and measurement error in both the alternative data source and the census.

To release public use files, further protections would be needed. Either a 20 per cent sample of records with age banded into five year groups (and a top category of 90+) or a 10 per cent sample with single year of age (and a top category of 75+) could plausibly be distributed as public use files. For a key consisting of seven variables there would be around 500 risky records in these samples that would be population unique according to the full set of census records. For a set of eight key variables there would be about 1400 risky records. Some of these are confounded with other cases in the sample with missing data on the key variables.

In the case of Trinidad and Tobago, a 10 per cent sample of records, including first level geographic codes (municipality) and single year of age, probably requires license use conditions for release. For keys of seven and eight variables, around 25,000 and 36,000 records out of 116,000 are population unique (according to the full set of census records). It should be emphasised that what these figures show is the number of individual records in this sample that could be re-identified by an intruder *that possessed the full census information for the key variables*. Of course, in reality intruders do not have this census information. They may have some external database with information about some members of the population, probably classified differently to the census variables, referring to a different point in time, with measurement errors and non-response. Even taking this into account, this is probably too many risky records to allow into a public use file.

For release as a public use file the sample could be reduced to a 5 per cent sample of records or age could be recoded into five year bands with similar effects on disclosure risk (Figure 2). Either of these files could be released as public use file with data swapping applied to mask the true geographic location of some risky households. Note though that it would not be possible to release both of these files (either based on the same or different samples) since disclosure risks cumulate with the release of multiple files.

III. Discussion: the dissemination of Caribbean census microdata

A. Demand for Caribbean census microdata

As already mentioned, Caribbean census microdata is not widely available to researchers. There are a relatively small group of countries and territories[1] that have disseminated anonymized census microdata to researchers. There are likely to have been cases of census datasets having been made available to researchers on an ad-hoc or informal basis.

This relatively limited dissemination of census microdata is easily understandable given the small size, and therefore limited human resources, of Caribbean statistical offices. For all statistical offices, developing procedures for the dissemination of microdata, and establishing and maintaining a responsive service which provides timely access to microdata, requires substantial resources. It is not only technically challenging but can also be complex in terms of the policy and legal issues involved.

It should also be acknowledged that just as the resources that can be devoted to microdata dissemination are limited, the number of researchers who are potential users of census microdata is also smaller than might be the case in larger countries. The Jamaican census data has been accessed by numerous researchers through the Derek Gordon Databank at the University of the West Indies. Applications to the Statistical Institute of Belize s Microdata Access Program and Suriname s General Bureau of Statistics Microdata Laboratory are growing from a low base with requests having come from both national users and international organizations. The Caribbean datasets available through IPUMS-International have not been well used although perhaps this is not surprising given that datasets for just two countries are available and the most recent census data is not available. For Jamaica the 1982, 1991 and 2001 censuses are available, and for Saint Lucia the 1991 and 1980 censuses, although it is anticipated that in the near future several Trinidad and Tobago censuses will be made available as part of the IPUMS-International series.

[1] Belize, Jamaica, Saint Lucia, Suriname and the United States Virgin Islands.

The limited use of official statistics microdata in research or policy analysis is partly due to difficulties in accessing the data (also the lack of data more generally). Statisticians can contribute to developing a culture of research and evidence based policymaking by extending access to microdata. If a country goes to the expense of carrying out a census, it is difficult to argue that resources are not available to disseminate or promote the use of the data. Indeed, the relatively limited number of statistical sources available is an argument for exploiting those that do exist as fully as possible.

To encourage greater use of microdata, researchers need to be made aware of its existence and the procedures for accessing it and they need to be provided with accurate and detailed metadata. Important initiatives in this regard are the International Household Survey Network s (IHSN) National Data Archive (NADA) and Data Documentation Initiative (DDI). NADA is a web-based cataloguing system that serves as a portal for researchers to browse, search, compare, apply for access, and download relevant census or survey information. The DDI is an international metadata standard designed to describe socioeconomic surveys, censuses, and other microdata collection activities. Both the NADA and DDI can help statistical offices to promote their microdata to a wider range of researchers.

B. The utility of small samples

A fundamental question concerning the anonymization of census microdata for small countries is at what point the technical disclosure controls applied to the datasets reduce their utility to such an extent that they are of little use. After their experiment anonymizing the 1991 census data for Saint Lucia, Levin and McCaa (2009) concluded that the technical confidentializing measures are so heavy that few researchers would wish to make use of this sample. The sample of anonymized records for the Saint Lucia census contains only 13,000 individuals which is certainly a small sample too small for the analysis of rare population characteristics such as health conditions, disability and recent births. The disclosure control measures applied to small census datasets certainly reduce data utility significantly and more than for larger countries.

However, the utility of samples of census data for small countries has to be seen in the context of the subregion where household surveys are infrequent and are themselves based on small sample sizes. For example, the most recent Survey of Living Conditions/Household Budget Survey (SLC/HBS) conducted in Saint Lucia in 2005/06 was conducted on the basis of a sample of 1,222 responding households (4,319 individuals) (Kairi Consultants Limited, 2007b). The Multiple Indicator Cluster Survey (MICS) conducted in 2012 was based on 1,718 responding households (Ministry of Social Transformation, 2014). A 10 per cent sample of anonymized census records from the 2010 census of Saint Lucia would probably contain about 4,200 households (16,500 individuals), significantly larger than other household surveys available. If it was possible to disseminate a 15 or 20 per cent sample of census records, then this would clearly offer to researchers a significantly larger sample of data than that available from household surveys.

To take the example of Grenada, the most recently available SLC/HBS (2007/08) was conducted based on a sample of 802 households (3,535 individuals). In the Labour Force Survey conducted in Grenada in 2014, 3,517 persons were interviewed (Government of Grenada, 2015). The 10 or 20 per cent samples of census records suggested above include 9,825 individuals or 19,240 individuals substantially more than anything available in a household survey.

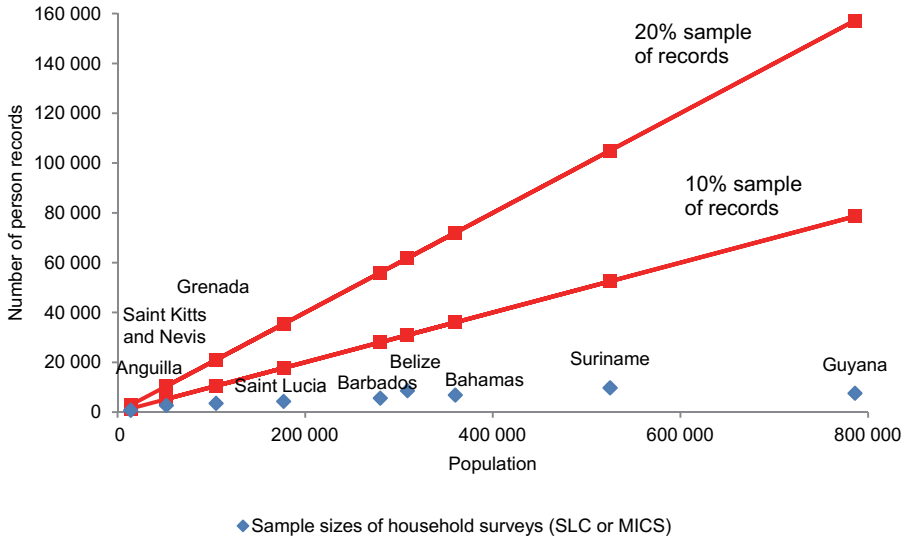
In Trinidad and Tobago, the Survey of Living Conditions (2005) was conducted based on 3,621 household interviews (12,919 person interviews) (Kairi Consultants Limited, 2007a). The MICS survey conducted in 2006 was based on 5,557 households containing 18,669 individuals (Ministry of Social Development, 2008). The 10 and 20 per cent samples of records considered here would comprise 116, 340 or 232, 300 individuals

Therefore, samples of census records still offer much larger samples than anything available from household surveys although the smaller the country, the smaller the differential between a sample of census records and a typical household survey.

This point is made clear in Figure 3 which compares the likely size of 10 and 20 per cent samples of census records with the sample sizes of recently carried out household surveys in those countries.

For countries with a population above 400,000, samples of census records still offer appreciably bigger samples than household surveys, whereas for smaller countries samples of records are barely any bigger than typical household surveys, even if the sampling fraction is increased to 20 per cent.

Figure 3
Sample sizes for 10 and 20 percent samples of census records compared with sample sizes of typical household surveys



Source: Economic Commission for Latin America and the Caribbean (ECLAC), on the basis of data provided by the Caribbean Development Bank (CDB) and The United Nations Children's Fund (UNICEF).

The problem of sample size can be overcome by releasing full count microdata, protected by other data reduction methods and licensed use conditions. This is not common practice among the mainly Anglophone countries whose methods have been reviewed as part of this study since it significantly increases the feasibility of re-identifying individual census records. However, it should be acknowledged that this assessment is based on a method of analysing disclosure risk which, although commonly applied, is known to provide a very conservative assessment of disclosure risk. Where full count microdata has been released under licensed use from Caribbean censuses, there have been no known breaches of confidentiality. What release of microdata in this way does mean is that the protection of confidentiality depends much more heavily on the administrative controls: the restriction of access to trusted researchers, the incentives for researchers to comply with the terms of data access agreements, and the threat of sanction in case of any breach of confidentiality.

C. Modes of dissemination

As mentioned above, statistical offices which have well developed procedures for the dissemination of microdata do not typically employ just one means of dissemination, for example only public use files. They will typically disseminate census microdata through a range of mechanisms. Using multiple means of dissemination allows statistical offices to maximise the research use of their data. Public use files can reach the widest range of users but necessarily contain the least detailed information. Licensed use files are only available to researchers, who must apply for access, a process which restricts access to a more select group.

The licensed use files themselves do however contain more detailed information than public use files. Data accessed on site in secure data laboratories are the most difficult to access, since the researcher has to visit the laboratory in person, but the reward for doing so is access to much more detailed microdata than is available in either a public use or a licensed use file.

The same is true for Caribbean countries. All Caribbean countries and territories are small, but Jamaica with a population of approximately 2.7 million is still 180 times larger than Anguilla with a population of around 15,000. The methods of dissemination appropriate for Jamaica are not likely to be appropriate for Anguilla. There is no single best method for disseminating Caribbean census microdata either for an individual country or for the subregion as whole. The different modes of dissemination all have a role to play. Some Caribbean countries could disseminate data through licensed use and public use files although clearly for smaller countries, samples of census records become too small to carry out useful analysis.

The alternatives to public use and license use files are to disseminate data either through remote access facilities or a secure data laboratory. In theory, remote access facilities could be a particularly fruitful form of data dissemination for small countries such as those in the Caribbean. They potentially offer controlled access to researchers to much more lightly protected microdata with much less suppression and recoding of variables and, crucially for small countries, without the need for sampling of records.

While Caribbean statistical offices themselves are not likely to have sufficient resources to devote to the development of such facilities, through partnership with data archives, Caribbean census microdata could be disseminated in this way. The Minnesota Population Center is working on the development of remote access facilities which in the near future could potentially provide controlled access to more detailed microdata than can be released through public use files or licensed use files. This could potentially provide researchers with the capacity to carry out statistical analysis of full count microdata.

The Caribbean has a ready-made solution for providing one form of remote access to census microdata (for tabulation and calculation of indicators) which is the REDATAM software. The REDATAM software was developed for storage, processing and dissemination of census data in Latin America and the Caribbean but it is currently underused in the Caribbean subregion. A major expansion in the use of REDATAM would be the most practical way of expanding remote access to Caribbean census microdata. The analysis of disclosure risk presented above can inform the way that REDATAM applications are designed for small countries so that they maximise the user s ability to remotely access microdata while still protecting confidentiality.

Secure data laboratories, such as those in Belize and Suriname, permit on-site access to detailed microdata but have two disadvantages. Firstly, they are not convenient to researchers who have to visit a particular location in order to conduct their analysis. As a consequence, the number of researchers using secure data laboratories tends to be relatively low. Secondly, they are relatively expensive to maintain. It was initially envisioned that the Derek Gordon Databank would operate in this way, with researchers visiting the databank to carry out analysis although it has not operated this way in practice.

An assessment of Derek Gordon Databank revealed that it could easily be adapted to function as a secure data laboratory. However, if the databank is to become a Caribbean data archive, serving the subregion as a whole, the limitations of a single on-site laboratory at the Mona campus of UWI in Kingston, Jamaica are obvious. More useful would be to focus on dissemination through public or licensed use files, and possibly in the future, some form of remote access if an interface could be developed to manage the submission and return of user queries.

IV. Conclusions

Caribbean countries face two specific challenges in the release of microdata which relate to the small size of the countries. Firstly, it is more difficult to disseminate analytically useful microdata in small countries because size itself provides some protection against the risk of disclosure. Secondly, the capacity of Caribbean statistical offices to devote human resources to statistical disclosure control and the dissemination of microdata is very limited. These are constraints which are not likely to change.

In respect of the technical constraint posed by small size, this issue becomes more acute the smaller the country. It influences choices about the mode of dissemination: public use, licensed use, remote access, or a secure laboratory. As discussed, there is generally no one single solution to the dissemination of microdata. All of these approaches to dissemination can play a role in opening up access to Caribbean census microdata.

Licensed or public use files which are samples of census records need to be of a minimally acceptable size. A sample of anonymized census records should offer something substantially bigger, and therefore more accurate, than a typical household survey would offer in that country. Samples of anonymized census records disseminated as either licensed or public use files can be analytically useful for the larger Caribbean countries (Jamaica, Trinidad and Tobago, Guyana and Suriname); they could be of some marginal utility for countries with populations of around 300,000 (Bahamas, Barbados and Belize); and are of limited value for countries with smaller populations.

Licensed use files are a relatively easy form of disseminating microdata. Licensed release of microdata offers an extra layer of protection compared to public use files and so is a sensible first step particularly for countries that have little or no experience of releasing census microdata. For a Director of Statistics to release a public use file of census records requires a lot of confidence in the disclosure risk analysis and the disclosure control methods. However, the example of the United States Virgin Islands demonstrates that it is possible even for very small countries.

REDATAM is underused in the Caribbean but has the potential to greatly extend access to census data. For other forms of remote access, for example offering researchers the ability to submit SAS, SPSS or STATA code to remotely held microdata, the Caribbean will most probably have to depend on developments outside the subregion, such as at Minnesota Population Center.

Important steps to open up access to microdata have been taken by the statistical offices of both Belize and Suriname, including release of census microdata in the form of licensed use files and through on-site access in secure laboratories. These are relatively recent initiatives which should be given time to establish themselves. In due course, the experiences of these two countries should provide valuable lessons for other statistical offices seeking to expand access to microdata.

Modernising statistical legislation to cover arrangements for microdata release would facilitate greater access to microdata by formally recognising this important function of statistical offices and providing a legal basis for data access agreements.

With respect to the capacities of Caribbean statistical offices, some realism about the human resources available to devote to statistical disclosure control is appropriate. Three levels of dissemination of census data were highlighted: published reports; online interactive tabulation; and dissemination of microdata (remote access facilities arguably have more in common with interactive tabulation tools than genuine dissemination of microdata). For Caribbean statistical offices, there is a clear order of priority attached to these three levels of dissemination. Published reports of tabulated data are the first priority for obvious reasons. Dissemination of census data through an interactive tabulation tool should be the second priority, with REDATAM being a proven and well-established mechanism for achieving this. Dissemination of microdata to researchers is the third priority. Given the demands on their time, and limited resources, many statistical offices are not likely to be able to devote significant resources to statistical disclosure control. Even much better resourced statistical offices than those in the Caribbean have found developing and applying methods for the dissemination of microdata to be difficult and frustrating. Many researchers have been frustrated either by the delays, the bureaucracy, or the unwillingness of statistical offices to release microdata.

Given the above, significant advances in the release of microdata will be most easily achievable through cooperation with data archives. Data archives are able to specialise in statistical disclosure control and in providing a service to researchers. To this extent, they serve as intermediaries between statistical offices and researchers and can, if not reduce the workload for statistical offices, at least enable them to expand access to microdata at relatively little cost. This can be seen as a service that the academic sector can provide to official statisticians. The *quid pro quo* is that statisticians have to allow access, in some form, to the data, first to the databank, and then through the databank to the researchers. The exact form that this access takes is negotiated and should be set down in written agreements between the statistical office and the data archive. Such written agreements set out the terms on which the databank can access the data, and their responsibilities to protect statistical confidentiality, and then the terms on which the databank can disseminate to researchers.

As previously mentioned, there are two data archives which have received and disseminated Caribbean census microdata: the Derek Gordon Databank based at the Mona campus of the University of the West Indies in Jamaica; and the international census archive (the IPUMS-International project) run by the Minnesota Population Center (MPC) at the University of Minnesota. There is substantial scope for expanding access to Caribbean census microdata through both these organisations.

The Minnesota Population Center s archive is the world's largest archive of publicly available census samples. They provide a platform for worldwide dissemination and have a wealth of experience in the technical, administrative and legal procedures to protect the confidentiality of census microdata. Statistical offices are encouraged to follow Jamaica, Saint Lucia and most recently Trinidad and Tobago in collaborating with the IPUMS-International project. The Minnesota Population Center plans to introduce remote access facilities to census microdata. This could be a particularly appropriate means of dissemination for smaller Caribbean countries which could help to overcome some of the limitations of public use and licensed files created from small census datasets.

Within the Caribbean, there is a lot of scope for strengthening the Derek Gordon Databank as a resource for Caribbean researchers. The activities of the IPUMS-International project and the Derek Gordon Databank can be of complimentary benefit to Caribbean researchers. The MPC disseminates on a worldwide scale, invests in harmonisation and integration of microdata for international comparability, and is well placed to develop infrastructure such as remote access facilities.

While the Derek Gordon Databank is not operating on the same scale as IPUMS-International, it could provide an expanded service to Caribbean researchers which could be enriched if more countries deposited their census datasets (and other datasets) with the databank. The dissemination of the Jamaican census data through the Derek Gordon Databank suggests that there is a demand for access to Caribbean census microdata and that a stronger local data archive, attuned to statistical environment in the Caribbean, would have no shortage of users. This would help to strengthen both academic social research and policy analysis in government.

ECLAC is already actively involved in promoting the widest possible use and dissemination of census data through its support for, and promotion of, the REDATAM software. Statistical disclosure control and confidentiality is therefore a natural area of interest. Through collaboration between statistical offices, international organisations, and data archives there is substantial scope for expanding access to Caribbean census microdata for researchers.

Bibliography

Cleveland, Lara and others (2012), 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 , Privacy in Statistical Databases 2012, Josep Domingo-Ferrer and Ilenia Tinnirello (eds.), Palermo, Sicily, September.

Domingo-Ferrer, Josep and Vicenç Torra (2008), A critique of k-anonymity and some of its enhancements *Availability, Reliability and Security* (ARES 08) Third International Conference.

_____(2001), Disclosure Control Methods and Information Loss for Microdata , *Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies*, Elsevier.

Dupriez, Olivier and Ernie Boyko (2010), Dissemination of Microdata files: Principles, Procedures and Practice , International Household Survey Network (IHSN), Working Paper No 0005, August.

Government of Grenada (2015), Findings of 2014 Labour Force Survey Revealed , [online], St. George s, Grenada [date of reference: 14 November 2015] <http://www.gov.gd/egov/news/2015/jun15/23_06_15/item_1/findings-2014-lfs-revealed.html>.

Harrison, Philomen (2012), Statistics and Open data , document presented at the Caribbean Open Data Conference and Code Sprint, Port of Spain-Kingston-Santo Domingo, January.

Hundepool, Anco and others (2010), Handbook on Statistical Control Version 1.2 , ESSNetSDC (A Network of Excellence in the European Statistical System in the field of Statistical Disclosure Control).

Ito, Shinsuke and Naomi Hoshino (2014) Data Swapping as a More Efficient Tool to Create Anonymized Census Microdata in Japan , Paper presented at Privacy in Statistical Databases 2014, Ibiza, Spain.

Kairi Consultants Limited (2009), Country Poverty Assessment, Grenada, Carriacou, and Petit Martinique, Technical and Statistical Appendices , Tunapuna, Trinidad and Tobago.

_____(2007a), Analysis of the Trinidad and Tobago Survey of Living Conditions, Tunapuna, Trinidad and Tobago, April.

_____(2007b), Trade Adjustment and Poverty in Saint Lucia 2005/06: Volume I: Main Report, Tunapuna, Trinidad and Tobago, June.

Levin, Michael and Robert McCaa (2009), Implementing IPUMS-International Confidentiality Protocols using CSPro/IMPS: 1991 Census Microdata of Saint Lucia , paper presented to the 17th Meeting of the Regional Census Coordinating Committee (RCCC, CARICOM), Castries, Saint Lucia, October.

Ministry of Social Transformation, Local Government and Community Empowerment and Central Statistics Office (2014), Saint Lucia Multiple Indicator Cluster Survey 2012: Final Report , Castries, Saint Lucia.

McCaa, Robert and others (2014), The IPUMS-International partnership enhances the value of census microdata for both producers and users , paper presented to the International Association for Official Statistics 2014 Conference on Official Statistics.

McCaa, Robert and others (2013), Analytical tests of controlled shuffling to protect statistical confidentiality and privacy of a ten per cent household sample of the 2011 census of Ireland for the IPUMS-International database , paper presented to Joint UNECE/Eurostat work session on statistical data confidentiality (Ottawa, Canada, 28-30 October 2013).

McCaa, Robert, Steven Ruggles and Matt Sobek (2010), IPUMS-International Statistical Disclosure Controls: 159 Census Microdata Samples in Dissemination, 100+ in Preparation , Privacy in Statistical Databases 2010, Josep Domingo-Ferrer and Emmanouil Magkos (eds.), Corfu, Greece, September.

McCaa, Robert and others (2006), IPUMS-International High Precision Population Census Microdata Samples: Balancing the Privacy Quality Trade-off by Means of Restricted Access Extracts , Privacy in Statistical Databases 2006, Josep Domingo-Ferrer and Luisa Franconi (eds.), Rome, December.

Ministry of Social Development (2008), Trinidad and Tobago, Multiple Indicator Cluster Survey 3, Final Report 2008 , Port of Spain, Trinidad and Tobago.

ONS (Office for National Statistics) (n/d), Statistical disclosure control for 2011 Census [online], Newport, United Kingdom [date of reference: 13 October 2015] <http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/statistical-methodology/statistical-disclosure-control-for-2011-census.pdf>.

Shlomo, Natalie (2010), Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility , *Journal of Privacy and Confidentiality*, vol. 2, issue 1.

Shlomo, Natalie, Caroline Tudor and Paul Groom (2010), Data Swapping for Protecting Census Tables , *Privacy in Statistical Databases*, *UNESCO Chair in Data Privacy, International Conference, Corfu, Greece, September 22-24, Proceedings*, vol. 6344, Domingo-Ferrer, Josep and E Magkos, Emmanouil (Eds.), Springer-Verlag Berlin Heidelberg.

Skinner, Chris and others (1994), Disclosure control for census microdata, *Journal of Official Statistics*, vol. 10, No. 1.

Templ, Matthias and others (2014), Introduction to Statistical Disclosure Control (SDC) , International Household Survey Network (IHSN) Working Paper No 007, August.

Templ, Matthias (2008), Statistical Disclosure Control for Microdata Using the R-Package sdcMicro , *Transactions On Data Privacy*, vol. 1, issue 2, Skövde, Sweden, August.

UK Data Service (2015), Census Microdata Guide [online], University of Essex, United Kingdom, [date of reference: 23 October 2015] <http://census.ukdataservice.ac.uk/use-data/guides/microdata>.

United Nations (2014), Fundamental Principles of Official Statistics , General Assembly resolution 68/261, March.

_____(2007), Managing Statistical Confidentiality & Microdata Access, Principles And Guidelines Of Good Practice , New York and Geneva, United Nations Sales No. E.07.II.E.7 ISBN 13: 987-92-1-116959-1 ISSN: 0069-8458.

United States Census Bureau (2015), United States Public Use Microdata Sample (PUMS), 2010 Census of Population and Housing, Technical Documentation , Washington D.C., May.

_____(2013), U.S. Virgin Islands Public Use Microdata Sample (PUMS), 2010 Census of Population and Housing, Technical Documentation , Washington D.C., September.

Annex

Annex

Technical notes

These technical notes describe the method that was used to analyse the disclosure risk associated with the release of samples of census records from the 2011 censuses of Grenada and Trinidad and Tobago. In commonly used notation, k=1,.....,K is used to denote the cells of the key, that is, all the possible combinations of values on the key variables (see for example Hundepool and others, 2010; Shlomo, 2010). The population size in cell k is denoted F_k so that:

$$\sum_{k=1}^{K} F_k = \text{N} \text{ where N is the total population}$$

Assuming that a sample of, in this case, census records is made available to researchers, the sample size in cell k is denoted f_k so that:

$$\sum_{k=1}^{K} f_k = \text{ n the total sample size}$$

For each key, k, there are a total of F_k records in the population which share the same values on the key variables and are therefore indistinguishable to the intruder. The individual risk of re-identification for record i can therefore be defined as:

$$r_k = 1/F_k$$

The individual record level risk is equal for all records in the same cell and it is for this reason that it is referred to as r_k instead of r_i. So for an individual record which is identical to 9 other records in respect of the key variables (so there are a total of 10 records sharing the same combination of values on those variables), the risk of re-identification for each of those records is 1/10 which is the probability of a successful match.

In many analyses of disclosure risk F_k is not observed. Where the dataset being released is a sample rather than a census, data for the population is not collected. Even in the case of census microdata, researchers may not be able to gain access to the full census dataset and so may carry out analysis on a sample of records. In these cases, r_k must be estimated. In the case of this study, full census datasets were made available and therefore the analyses are based on both the full set of census records and samples of records selected according to the methodology proposed here for creating microdata release files. Therefore F_k is observed and r_k can be calculated directly.

Of course, no census enumerates the population completely and each of the datasets analysed contains non-response weights to adjust for unit non-response. These weights imply non-response rates of around 10 per cent in Grenada, and 12 per cent in Trinidad and Tobago. Unit non-response implies that the F_k s will be underestimated and therefore the r_k s over-estimated. So all the measures of risk presented here are conservative for this reason. However, this over-estimation of risk is only of the order of a few per cent. A unit non-response rate of 10 per cent does not imply that risk is over-estimated by 10 per cent but by something much less than that.

It is assumed that an intruder attempts to re-identify individuals rather than households since re-identification of individuals through variables such as age, marital status, occupation, and place of birth is a greater threat than re-identification of households through household variables containing information about household type, structure, or other household characteristics. External databases typically contain information about individuals rather than households (although a nosy neighbour is more likely to have knowledge of household structure). The primary focus here is therefore on the risk that individuals are re-identified. However, the risk of household re-identification can also be calculated. A household is considered to have been re-identified when one of the individuals in the household has been identified. The risk of re-identification of a household r_h can therefore be calculated from the

product of the complements of the individual risks for the members of the household (since the risk of re-identification of a household is the complement of the probability that none of the individuals in the household are re-identified):

$$r_h = 1 - \prod_{i \in h} (1 - \frac{1}{F_k})$$

This measure of household risk is used to target the data swapping algorithm towards high risk households which are selected with probability proportional to the risk r_h.

These individual (or household) level measures of risk can be combined into global measures of risk for a whole file of microdata records. A range of measures of risk are calculated for different microdata release files in order to illustrate how a release file can be designed to minimise disclosure risk.

The global risk R is the sum of the individual record risks for all the records in the release file

$$R = \sum_{i=1}^{n} \frac{1}{F_k} \text{ or alternatively } \sum_{k=1}^{K} \frac{f_k}{F_k}$$

and can be interpreted as the expected number of correct matches should an intruder attempt to match the microdata release file to an external database using a given key. This measure provides a general indication of the level of risk associated with different microdata release files although there is an important distinction to be drawn between the expected number of correct matches and the number of unique matches. Only when a record is unique on the key variables can an intruder be sure that they have identified a record. It is when the intruder is able to verify such a link that disclosure is regarded as having taken place (Skinner and others, 1994).

Sample uniques, records which are unique in the sample for a given key (that is $f_k = 1$) are generally at higher risk than records which are not unique in the sample (records for which $f_k > 1$). The total number of sample uniques, SU, is given by:

$$\text{SU} = \sum_{k=1}^{K} I(f_k = 1) \text{ where I} = \begin{cases} 1 \text{ if } f_k = 1 \\ 0 \text{ otherwise} \end{cases}$$

And the expected number of correct matches for sample uniques:

$$\text{E(matches for SU)} = \sum_{k=1}^{K} I(f_k = 1) \frac{1}{F_k}$$

both of which can also be expressed as a proportion dividing by the sample size n.

Records which are merely unique in the sample but not in the population are protected by the fact that while an intruder can distinguish them from all other records in the sample, these cases are indistinguishable from others in the population so the intruder cannot match uniquely to a single individual in the population. Population uniques are the records most vulnerable to re-identification since an intruder possessing information for the key variables for the population could uniquely identify these records in the release file. PU, the number of population uniques in the sample, and K3, the number of records which are k-anonymous (k=3), are important indicators of disclosure risk.

$$\text{PU} = \sum_{k=1}^{K} I(F_k = 1)$$

$$K3 = \sum_{i=1}^{n} I(F_k < 3)$$

This framework has been used to assess disclosure risk for a range of different designs of microdata release files for a range of different keys. However, it has been extended to account for the presence of missing data (item non-response). Most analyses of disclosure risk do not take account of the impact of missing data on key variables. Where statistical offices have applied methods of imputation to replace missing values on key variables, this is not a problem (or at least it is a different problem: how to take account of imputation in an assessment of disclosure risk). However, such techniques are not routinely applied to Caribbean census datasets and for those analysed, a majority of key variables included some missing values. Missing values are those coded as not stated rather than those coded as not applicable . For example, in the case of the variable occupation, the value for a child, for whom the concept of occupation is not applicable, is not considered missing but the value for an adult who doesn t provide a response to the question, would be considered missing.

Missing data is generally assigned a code, for example -9. It is clear that in an analysis of disclosure risk, the value -9 cannot be treated in the same way as other categories because it contains no information about the individual (except that they didn t provide a response to that census question). The presence of missing data for key variables would thus hinder the efforts of an intruder to re-identify individuals. Treating missing data as if it were a category like any other leads to an over-estimation of disclosure risk.

The analysis carried out here was run initially treating missing values as if they were simply another category, but then two adjustments were made in order to account more properly for missingness in the key variables. Here, a distinction is made between the cells which include at least one missing value code for at least one key variable, denoted k'', and all other cells which consist entirely of valid responses (including not applicable) which are denoted k'. In the presence of missing data, the cells can no longer be considered to partition the population and sample in a mutually exclusive way with each individual belonging to one and only one cell of the key. The cells k'' only arise because data is missing and the records in these cells really belong in one of the cells k', it is just that the cell to which these records belong is unknown. For a given cell k'' containing one or more missings, let $S(k'')$ be the set of cells k' to which the records in cell k'' could possibly belong.

The uncertainty created by missing data reduces disclosure risk. The records with missing data clearly have lower record level disclosure risk than if the data wasn t missing (or if missing values are treated like any other valid response) since an intruder has less information about these individuals with which to identify them. In addition, the individual disclosure risk for records without missing data is also lower since these records can be confounded with records with missing data. A record with no missingness might appear to be unique but there may be records with missing values which could be identical to them on the key variables.

This issue is handled here by adjusting the sample and population cell counts to reflect the uncertainties associated with missing data and calculating an adjusted measure of individual record level risk accordingly. For cells k'' containing one or more missings, all records in cells $k' \in S(k'')$ are indistinguishable from the records in cell k''. To reflect this, adjusted sample and population cell counts ${F^*}_{k''}$ (and ${f^*}_{k''}$) and adjusted measures of individual risk ${r^*}_{k''}$ are calculated in the following way:

$$ {r^*}_{k''} = 1/{F^*}_{k''} \quad where \quad {F^*}_{k''} = F_{k''} + \sum_{k' \in S(k'')} F_{k'} $$

Note that $S(k'')$, the set of cells k' to which the records in cell k'' could belong, are those cells which result from replacing the missing values with every possible combination of valid responses. Valid responses in this case do not include not applicable . It is assumed that if data is missing, that is not stated , then the true value for the missing item cannot be not applicable (for the value to be not stated , the question must have been asked, and therefore the question must have been applicable to the respondent).

For records with missing data which are included in the sample of records, an equivalent adjustment is made to the sample cell counts and quantities $f^*_{k''}$ are calculated as follows:

$$f^*_{k''} = f_{k''} + \sum_{k' \in S(k'')} f_{k'}$$

In the case of records with keys k' without missing data which are included in the sample, a slightly different kind of adjustment is made to reflect the way in which the presence of missing data in records in the sample (rather than the population) can confound the other records in the sample in which there is no missingness. This adjustment effectively takes the number of records in each cell k'', and reapportions them across all the cells $k' \in S(k'')$ in which they could possibly belong. These records are distributed in proportion to the number of sample records already in the cells k'. Note that in general this apportionment gives rise to quantities which are fractional, mainly between 0 and 1, and therefore whereas F_k is a whole number, this will not necessarily be true of F^*_k. Let $T(k')$ be the set of cells k'' containing records which could possibly belong in cell k' depending on unknown missing data. More formally:

$$r^*_{k'} = 1/F^*_{k'} \ where \ F^*_{k'} = F_{k'} + \sum_{k'' \in T(k')} f_{k'} \frac{f_{k''}}{\sum_{k' \in S(k'')} f_{k'}}$$

The corresponding adjustment of the sample cell size is:

$$f^*_{k'} = f_{k'} + \sum_{k'' \in T(k')} f_{k'} \frac{f_{k''}}{\sum_{k' \in S(k'')} f_{k'}}$$

Together these adjustments account for how the presence of item non-response in key variables reduces the risk of disclosure. The calculation of the quantities $r^*_{k''}$, $F^*_{k''}$ and $f^*_{k''}$ reflect the lower risk of re-identification for records with missingness in key variables. The calculation of the quantities $r^*_{k'}, F^*_{k'}$ and $f^*_{k'}$ accounts for the confounding effect of sample records with missingness upon sample records without missing values. Sample records with missingness can also confound other sample records with missingness which has not been addressed here, however, this is a second order effect whose impact on aggregate measures of disclosure risk is likely to be small.

So to summarise:

$$F^*_k = \begin{cases} F_{k''} + \sum_{k' \in S(k'')} F_{k'} & for\ keys\ with\ missing\ data \\ F_{k'} + \sum_{k'' \in T(k')} f_{k'} \frac{f_{k''}}{\sum_{k' \in S(k'')} f_{k'}} & for\ keys\ without\ missing\ data \end{cases}$$

with f^*_k defined in a similar way.

It remains only to clarify how these adjustments have been applied in the calculation of the various aggregate measures of risk presented for the different microdata release file designs. The second adjustment described above produces F^*_k s and f^*_k s which are fractional with many just slightly larger than one. These correspond to records which are probably either sample or population unique, but might not be unique depending on the unknown missing values taken by other sample records. In this analysis, and in the tables above, these near uniques are still counted as sample or population uniques, but a separate count of the near uniques is provided described as uniques which are confounded by other cases with missing data on the key variables. These cases can be considered as introducing uncertainty to any claim an intruder may make to have identified an individual. Recognising the number of sample and population uniques which are near uniques can usefully inform decisions about the number and proportion of uniques that can satisfactorily remain in a microdata release file.

Therefore the numbers of sample uniques are calculated based on the adjusted $f^*_{k''}$ s for the cells with missing data but on the original $f_{k'}$ s for the cells without missing data:

$$\text{SU} = \sum_{k'} I(f_{k'} = 1) + \sum_{k''} I(f^*_{k''} = 1)$$

The count of the near uniques , those records which are confounded with other sample records with missingness, is the difference between this figure and that which would have been calculated had adjusted $f^*_{k'}$ s also been used for records with no missingness. The number of population uniques is counted in the same way:

$$\text{PU} = \sum_{k'} I(F_{k'} = 1) + \sum_{k''} I(F^*_{k''} = 1)$$

The number of sample records which are not k-anonymous in the population for k=3 is equal to:

$$K3 = \sum_{i=1}^{n} I(F^*_k < 3)$$

The expected number of correct matches for sample uniques is:

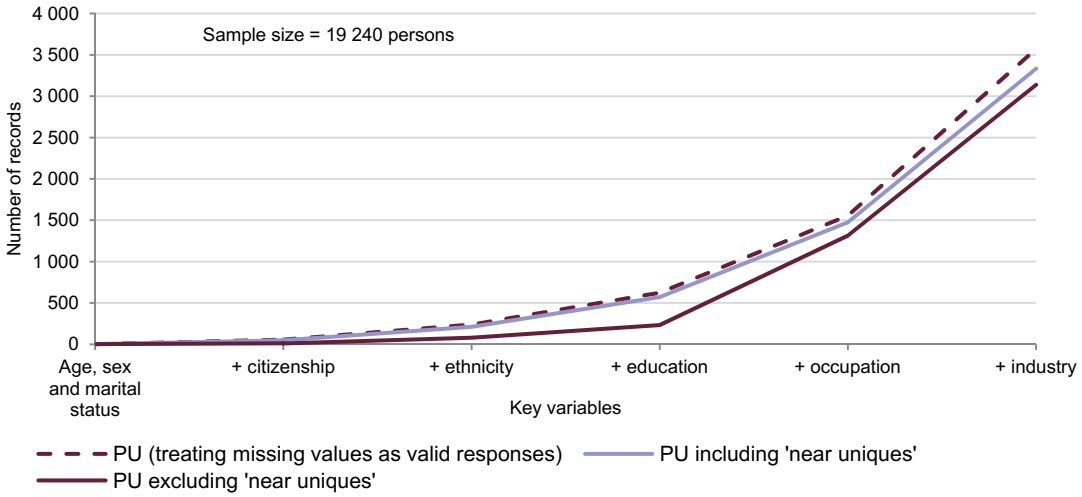
$$\text{E(matches for SU)} = \sum_{i=n}^{n} I(f^*_k = 1)\ \frac{1}{F^*_k}$$

The global risk for the whole file is the sum of the individual record level risks:

$$R = \sum_{i=1}^{n} \frac{1}{F^*_k}$$

Figure A.1 shows the effect of these two different adjustments and how taking account of missing data on key variables produces a lower estimate of disclosure risk. The confounding effect of sample records with missing values on attempts at re-identifying records is particularly significant for shorter and mid-length keys for which a significant proportion of uniques are confounded with other sample records with missing data. For these cases, it would be impossible for an intruder to claim with certainty that they had re-identified an individual.

Figure A.1
The number of population uniques in a 20 percent sample of records from the 2011 Census of Grenada calculated using different treatments of missing data



Source: Economic Commission for Latin America and the Caribbean (ECLAC), on the basis of data provided by the Central Statistical Office of Grenada.

UNITED NATIONS

ECLAC **Studies and Perspectives – The Caribbean**

Issues published

A complete list as well as pdf files are available at

www.eclac.org/publicaciones

49. Dissemination of Caribbean census microdata to researchers - Including an experiment in the anonymization of microdata for Grenada and Trinidad and Tobago, LC/L.4134, LC/CAR/L.486, 2016.
48. An assessment of big data for official statistics in the Caribbean Challenges and opportunities, LC/L.4133, LC/CAR/L.485, 2016.
47. Regional approaches to e-government initiatives in the Caribbean, LC/L.4132, LC/CAR/L.483, 2016.
46. Opportunities and risks associated with the advent of digital currency in the Caribbean, LC/L.4131, LC/CAR/L.482, 2016.
45. Ageing in the Caribbean and the human rights of older persons Twin imperatives for action, LC/L.4130, LC/CAR/L.481, 2016.
44. Towards a demand model for maritime passenger transportation in the Caribbean A regional study of passenger ferry services, LC/L.4122, LC/CAR/L.477, 2015.
43. The Caribbean and the post-2015 development agenda, LC/L.4098, LC/CAR/L.472, 2015.
42. Caribbean synthesis review and appraisal report on the implementation of the Beijing Declaration and Platform for Action, LC/L.4087, LC/CAR/L.470, 2015.
41. An assessment of performance of CARICOM extraregional trade agreements An initial scoping exercise, LC/L.3944/Rev.1, LC/CAR/L.455/Rev.1, 2015.
40. Caribbean Development Report Exploring strategies for sustainable growth and development in Caribbean small island States, LC/L.3918, LC/CAR/L.451, 2014.
39. Economic Survey of the Caribbean 2014 Reduced downside risks and better prospects for recovery, LC/L.3917, LC/CAR/L.450, 2014.
38. An assessment of mechanism to improve energy efficiency in the transport sector in Grenada, Saint Lucia, and Saint Vincent and the Grenadines, LC/L.3915, LC/CAR/L.449, 2014.
37. Regional integration in the Caribbean The role of trade agreements and structural transformation, LC/L.3916, LC/CAR/L.448, 2014.
36. Strategies to overcome barriers to the implementation of the Barbados Programme of Action and the Mauritius Strategy in the Caribbean, LC/L. 3831, LC/CAR/L.441, 2014.
35. Foreign direct investment in the Caribbean Trends, determinants and policies, LC/CAR/L.433, 2014.

STUDIES AND PERSPECTIVES

49

STUDIES AND PERSPECTIVES

ECLAC Series

ECONOMIC COMMISSION FOR LATIN AMERICA AND THE CARIBBEAN
COMISIÓN ECONÓMICA PARA AMÉRICA LATINA Y EL CARIBE
www.eclac.org