

PRASC



**Project for the Regional
Advancement of Statistics
in the Caribbean**

**Projet régional pour
l'avancement de la statistique
dans les Caraïbes**

Funded by the
Government
of Canada

Canada



Panel presentation – Perspectives on statistical confidentiality for census tables

Project for the Regional Advancement of Statistics in the Caribbean (PRASC)

Jean-Louis Tambay, Statistics Canada
November 10, 2020



Delivering insight through data for a better Canada

A Sacred Trust

- Statistical Agencies make a solemn promise to respondents that they will safeguard the confidentiality of information provided to them.
- Possible consequences of failure to do so:
 - Distress or harm to respondents.
 - Lawsuits.
 - Loss of trust in the Agency, leading to a drop in respondent collaboration, an increase in collection and processing costs, and lowered confidence in the Agency's statistics.

Disclaimer: The views expressed in this presentation are those of the presenter alone and do not necessarily represent those of Statistics Canada or the Government of Canada.

Disclosure from Census Frequency Tables

- Accidental/unintentional
 - Spontaneous recognition of a public figure, relative, friend, acquaintance ... more likely within small sub-populations
 - Attribute disclosure about an identified individual, or a group, e.g., narrow income range for members of a certain profession
 - Self-identification (1-respondent cell) leading to a public complaint
- Intentional
 - Opportunistic: Search data for *low hanging fruit* (e.g., counts of 1) by a privacy advocate or investigative journalist, or for recognition
 - Targeted: Seeking info about a public figure (e.g., senior government official)
 - Reconstruction attempts: Trying to gain information about persons on a database or a census public microdata file

DC Strategies Considered for Frequency Data

- Pre-tabular: Microdata perturbation, e.g., swapping
 - Can be difficult; impact may be too severe, or not enough
- Post-tabular: Additive noise, Random rounding
 - Overall a more efficient way to protect census data
 - RR is a special case of AN, thus it offers less flexibility
 - Rounding more “visible”, e.g., RR: **6 + 0 + 3 = 12**; AN: **7 + 1 + 3 = 9**
 - Smaller noises and rounding bases offer less protection
 - Consistency (cellKey) impedes repeated query attacks, but can reveal “false zeroes”
- Table restrictions
 - Disallowing queries, e.g., multi-dimensional table with detailed geography
 - Suppressing query outcomes, e.g., sparse tables (low mean cell size)
 - May be too severe, which is where other approaches come in

Risks with On-Line Query Systems

- Increased likelihood of accidental disclosure
- Opportunities for hackers to target individuals through differencing and by exploiting inconsistencies in results
- Repeated requests can undermine the randomization process (averaging attacks)
- Algorithms can tighten ranges in “noisy” counts and produce ranges for higher dimensional results by linking lower dimensional tables
- Outputs can be linked to public/private datafiles to gain information on units
- Restrictions such as minimum population sizes can be circumvented
(e.g., $\#FemaleCEOs = \#All - \#Males - \#NonCEOs + \#MaleNonCEOs$)
- User-generated Derived Variables (e.g., custom areas or income ranges, variables combining multiple attributes) can be created to bypass some table restrictions

(Accidental) attribute disclosure

- Consider the following perturbed results for a Region A

Education	Incomplete	High School	College	University	All
Ethnicity					
• Chinese	0	0	3	0	2

- If there is only one Chinese household known to live in the region, we have disclosed that some members have a college degree.
- Problem: Uniqueness plus local knowledge (i.e., certainty of no other Chinese households in the region)
- Solution 1: Provide less detail on Ethnicity, especially at the regional level
- Solution 2: Suppress counts for visible characteristics like Ethnicity=X when they relate to only one household (or two) in the region

Targeting known individuals

- If it is known that only one politician in a region is under 30 years old

SCENARIO A	Univ. Degree [YES]	Univ. Degree [NO]
Occup. = Politician	32	9
• Age over 30	32	7
• Age under 30	0	3

SCENARIO B	Univ. Degree [YES]	Univ. Degree [NO]
Occup. = Politician	32	9
• Age over 30	32	8
• Age under 30	0	0

- Both scenarios reveal the absence of a university degree for that person
 - The count of 3 implies the true count is not zero (thus it is 1)
 - Inconsistent perturbed values imply the 0 is not a true zero ($9 \neq 8 \Rightarrow 0 \neq \text{true } 0$)
- Problem: Known uniqueness, plus ability to exploit it
- Solution 1: Better control of user queries (e.g., less detailed occupation)
- Solution 2: Protect vulnerable targets by modifying their microdata (e.g., age)

Importance of a multi-level approach

- Post-tabular methods are quite efficient, but not infallible
- It is necessary to supplement them with other approaches
- **Query restriction:** esp. below national level (limit # dimensions, types & combinations of variables, level of detail, ...)
 - Manage expectations: System may not meet all user demands, consider other forms of access, e.g., custom requests
 - Manage requests: Limit ability to create user defined variables
- **Microdata perturbation:** the most vulnerable units can be identified and masked (e.g., change age/occupation/ethnicity) or excluded from on-line queries
- Query monitoring can go a long way in preventing and discouraging attacks

Importance of a holistic approach

- Consider problem in its entirety: data, users, uses and outputs (the “5 safes”)
 - More control on users allows less severe restrictions on outputs
 - Provide alternate access for analyses not possible under system
 - Err on the side of caution: it is easier to relax restrictions than to tighten them
- Be mindful of risks from multiple data products:
 - Census tables can increase information content of census public microdata files
 - Tables of magnitude data have different problems and solutions
 - Releasing means and totals can affect the protection of counts (& vice versa)
 - Tables of distribution (e.g., for age or income ranges) present additional risks, especially if ranges can be manipulated
 - Some analytical outputs can reproduce tabular results

You can contact the PRASC team at:
statcan.prasc-prasc.statcan@canada.ca