Statistics Canada
www.statcan.gc.ca

Development of rules from administrative data

Michelle Simard

Statistics Canada

UNECE Worksessions on Statistical Disclosure Control Methods

Helsinki, October 2015

Outline

- Progress Status - Canada
- Administrative Data
- Proposed solutions
  - for risk: the Companion
  - for method: The Layered Perturbation Method
- Future

Current Status

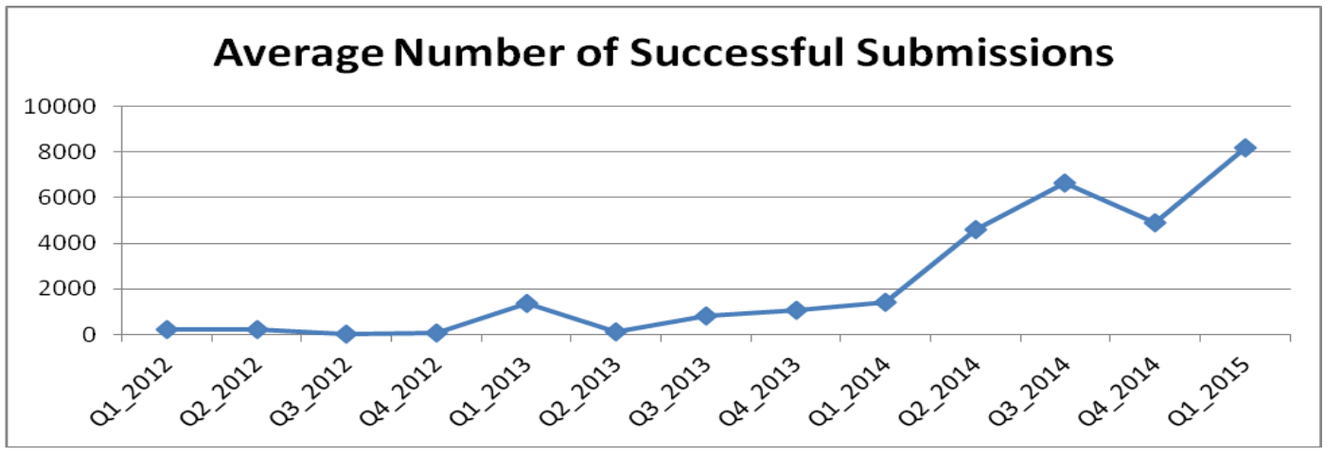
- **G- CONFID**
  - For skewed data - mostly business surveys
  - Primary (PCS) and Complementary Cell suppression (CCS)

  - New in 2015:
    - Additive controlled rounding (G-TAB)

- For Fall 2016, the automated treatment of:
  - Data in the presence of waivers
  - Variables containing both positive and negative values
    - Absolute values
    - Linear combinations including proxy variables
  - Survey weights

Current Status – Social Surveys

- **Real Time Remote Access (RTRA)**
  - Started in 2009
  - Currently 29 social surveys (+ cycles), 5 administrative files
  - International Users



- **Generalized Tabulation System (G-TAB)**
  - Integration with our common tools infrastructure, generalized systems and our new dissemination model
  - GUI, Engine and Confidentiality modules
  - Starting transition of social surveys to G-TAB
- Both tools share the same Engine and Confidentiality modules

Current Status – G-TAB

- Created efficiencies, robustness and responsiveness in our process
- Improved quality and standardization (rules, methods, metadata)
- Accelerated data availability
- SDC methods developed for
  - Counts, means, percentiles, quartiles, proportions, ratios, Gini coefficient, geometric means, level change, percentage change, moving averages, Z-tests
- Additive and Controlled rounding
  - Other rules : Min cells rule, no 0 and no 1 rules, etc…
- Quality indicators (C.V., S.E)
- What's next ?

  Administrative data and TOTAL

Administrative Data

- **Why not treating them like survey**
  - Census of a given population
  - No sampling, no weights
  - Outside StatCan
  - Different ownership, governance and policy framework (STC)
    - Often acquired under a legal framework from known institutions
      - Own dissemination approach
    - Release is still under the Statistics Act
- Continuous increase of administrative data in all programs
- Our dissemination activities and our business evolved
  - More complex than what it used to be internally
  - In the past 3 years : defined or created new
    - division, policies, governance, ownership framework …

Administrative Data: some definitions

- **Type A:**
  - Discrete variables or continuous variables with no dominance
  - Institutions-type files
- **Type B:**
  - Continuous variables, fiscal-type information, skewed distribution
  - Taxation files, immigration database
- **Other types:**
  - Census linked with one administrative file
  - Linkage of administrative files and/or survey files

- Dealt with types A and B

Administrative Data - Balancing Act

- Develop options considering also
  - Users (internal vs. external)
  - Types
  - Governance

- Impossible to develop a “no risk” approach
- Different types, different ways of using them
  - Risk management has to be integrated
  - Needed a framework more than just methods

Administrative Data – The Approach

- Disclosure control framework - 3 R: the rules, the risks, the roles

  - Risk: the Companion
    - Support, guide, help and TEACH the analysts, users, directors about the disclosure risk
  - SDC rules
    - Introducing the Layered Perturbation Method (LPM)

  - Roles and responsibilities
    - Confidentiality methodologists GTAB/RTRA
    - Administrative data (data sources) methodologists
    - Administrative data (data sources*)* analysts
    - Directors (owner of the acquired files)

The Companion

- Disclosure risk management tool
  - To be implemented for various systems, outputs and tools
  - Provides decision-support information
  - Does not replace the expertise
- Plans for G-TAB Companion
  - Scan of the outputs (tabular, statistical analyses)
  - Provide a quick assessment on how sensitive a table is (with *a global score* or a **color-code**)
  - Provide a detailed log of the potential disclosure risks of a table or model
  - Suggest possible solutions
- Users can then decide to
  - Release or not the table as is
  - Take appropriate measures to ensure that the risk is more manageable

The Companion

- **The risk is measured based on:**

1. Geographic levels
   - National
   - Provincial
   - Sub-provincial – High level
   - Sub-provincial – Low level
2. Sensitive values
   (ex: Rare type of cancer)
3. Proportion of cells with low counts(1 or 2)
4. Presence of full cells
5. Presence of dominance (Type B : means and totals)
6. Presence of derived variables

Requires client's input

Automated process

Score Function

The Layered Perturbation Method

- Developed for personal taxation data in a custom tabulation environment
- Few dominant units in a cell total
- Covers residual disclosure from multiple tabular requests (focus on differencing)

- Benefits
  - Protection of ratios
  - Treatment of zeroes and negative values
  - Maintenance of data quality (minimal loss information)
  - Minimal manual intervention
  - Computational simplicity

The Layered Perturbation Method

- Suppress sensitive cells only (no CCS)
- Perturb units in all other cells (e.g., using Evans-Zayatz-Slanta (EZS) multiplicative noise $w_i=1+\varepsilon_i$ for $\varepsilon_i \sim (0,\sigma_\varepsilon^2)$)
- Largest units perturbed consistently
- Median units perturbed semi-consistently
- Smallest units not perturbed

The Layered Perturbation Method

- Basic idea 1: Pseudo-random hash numbers (h_i)
  - Attached to each unit
  - E.g., h_i ~ Uniform(0,1) used to determine unit-specific noise w_i
  - h_i' used to determine unit-cell-specific noise w_i'

- Basic idea 2: *Layered* perturbation
  - Largest n_1 units are always perturbed consistently using w_i
  - Next n_2 units are perturbed semi-consistently
    - Use a mixture of w_i & w_i' for those n_2 units (unit-specific and cell-specific)
  - Remaining units are not perturbed (no noise)

  - Similar to ABS TableBuilder

The Layered Perturbation Method

- Basic idea 3: Focus on differencing problem
  - Increase noise for top 3 units, if needed
  - Set w_i from $(-1)^{i+1}\varepsilon_i$ to increase variance of differences
  - Noises change direction if a top contributor is removed (even-ranked)
  - Lessens risk when small unit is added/removed

- Suppress sensitive (e.g. p%-rule) and small cells (e.g. n<15)
- Perturb largest n_1+n_2 units in other cells following the method

Evans, T., Zayatz, L. and Slanta, J. (1998). Using Noise for Disclosure Limitation of Establishment Tabular Data. *Journal of Official Statistics,* ***14, 537–551.***

Thompson, G., Broadfoot, S. and Elazar, D. (2013). Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics. *Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa, 28-30 October, 2013.*

Future

- Social data issues
  - Residual Disclosure
    - Monitoring the situation for social surveys
  - Increasing usage of Administrative Data
    - Big data
    - Linkages
    - Governance/Approval protocols before releasing

- Census
  - Reengineering their tabulation system