Marrying demand for statistical information with disclosure control: The Canadian experience in developing an automated dissemination tool in an open-data world

March 9, 2018

Zixin Nie (Statistics Canada)
Claude Girard (Statistics Canada)
Outline

• Evolution of Data Access
• The Generalized Tabulation system (GTAB)
• Results and the current state of GTAB
Evolution of Data Access

- Statistics Canada has a mandate to gather data and report upon the findings to all Canadians
- Main vehicles of dissemination are the Statistics Canada Daily Report (*The Daily*) and CANSIM (system for viewing official tables)
  - Published products of aggregate statistics
  - Relatively high-level overviews
Evolution of Data Access

- Main modes of directly accessing Statistics Canada microdata by external users are Public Use Microdata Files (PUMFs) and Research Data Centres (RDCs).
- PUMFs are modified microdata files that minimise the risk of disclosure of confidential information.
- RDCs give users direct access to unmodified STATCAN microdata, but require special permission to access.
Creation of the Generalized Tabulation System

• Generalized Tabulation System (GTAB), was borne from new needs arising from greater demand for data access
What is the GTAB system?

• Generalized System
• Next generation tabulation and dissemination tool
• Corporate tabulation tool
  • Social, health, and labour statistics
• Direct pipeline from microdata to publishable tables
Broader vision of GTAB

• Standardization of practices within Statistics Canada
  • GTAB not a system designed to replicate all functionality of previous systems
  • Move towards more standard practices across surveys in microdata structure, estimation, dissemination, confidentiality
  • Main benefit: making published products easier to comprehend through similar structure
  • Another benefit: skills obtained through working on one survey can be easily transferred to working in other areas
Broader vision of GTAB

• Create an easy to use system for users without programming experience
  • Other systems used at STATCAN require experience with coding in SAS
  • GTAB needs to be accessible to users who do not have extensive coding experience
  • GUI – simple enough for most users to learn and use quickly
The GTAB Framework

- Generalized process to dissemination
- Streamlined pipeline from microdata to publishable tables
  - Takes final microdata files as input (with survey weights and replicate weights)
  - Create table specifications
    - i.e. Select domain variables to cross, statistics to be calculated
  - Outputs can directly be disseminated

Single Microdata File

Calculation Engine
- Calculates statistics
- Calculates precision measures

Assign Quality Indicators based on Precision Measures

Apply Confidentiality

Create final output for dissemination
The GTAB Framework

- GTAB will automatically
  - Calculate precision measures using replicate weights (Rao-Wu-Yue bootstrap weights)
  - Assign standardized quality indicators based on coefficient of variation
  - Apply confidentiality rules
- Data flow is linear, we do not pass information back to previous steps
- Cannot combine multiple files for calculation within GTAB
The image shows a screenshot of a user interface for a statistical data tabulation tool. The tool is used for creating tables and analyzing statistical data. The interface includes options for selecting variables and adding different types of statistics. The tool is part of the GTAB - Tabulation Tool (GRID TEST Environment) which is used for testing survey methodologies. The interface shows various options for adding and removing variables, selecting domain variables, and specifying statistic specifications. The tool also includes options for adding different types of statistics such as weighted-frequency, percent-distribution, and higher-order statistics. The interface displays messages indicating the status of the tool and any connection issues.
GTAB Functionality Development Process

- Data providers (clients) approach GTAB team for dissemination
- Demand VS supply assessment: Clients’ needs VS GTAB’s functionalities
- Standardization: Significant business case must be made before turning yet-unfulfilled needs into new system specifications
GTAB Functionality

• Statistics that GTAB can currently calculate
  • Level 1 statistics: Mean, percentiles, median, total, weighted frequency
  • Level 2 statistics: Gini coefficient
  • Level 3 statistics: Proportions and ratios
  • Level 4 statistics: Moving averages over time
  • Level 5 statistics: Level change, percentage change, significance tests (Global, base value, sequential, sequential over time)
  • Quantiles, both as domain variables and as bound statistics
  • All calculated statistics use survey weights

• Precision measures
  • Variance, standard error, coefficients of variation, confidence interval bounds
GTAB Functionality

• Confidentiality rules
  • Each statistic currently available in GTAB has its own set of confidentiality rules
  • Rules are applied equally, regardless of subject matter
  • Tested through simulation studies on fake and real data, vetted by experts, and approved through management
  • ACRound, suppression based on minimum counts, rounding of final outputs
  • Parameter-driven
  • Rules are automatically applied to outputs when requested
Current state of GTAB

• GTAB can calculate ~90% of statistics found in published tables for social, health, and labour statistics

• Numerous surveys are transitioning their dissemination to GTAB, such as Canadian Community Health Survey, Education Surveys, Tourism and Travel surveys, and Labour Force Survey

• Census moving dissemination to GTAB

• New functionality in constant development to meet new business requirements
Advantages to using GTAB

• Dissemination becoming more standardized
  • Tables from many surveys now use standard confidentiality rules, standard quality indicators, and standard methods for calculating statistics and precision measures
  • Skills obtained when disseminating for one survey are now useful for many different surveys
  • Users of published STATCAN data on CANSIM now have information presented in a more uniform fashion, increases usability of data
• Promotion of improved methods for reporting quality of estimates, such as publication of confidence intervals
• Promotes better practices internally when creating pre-dissemination files
Advantages to using GTAB

• Provides precision estimates for a variety of statistics using replicate methods (bootstrapping)
• Increased timeliness for custom tabulations
• Engine used in GTAB system is also being used to power other systems for automated dissemination
Conclusion

• Demands from users of Statistics Canada data drove need to develop new system for dissemination and confidentiality
• Developed GTAB as an easy-to-use tabulation system
  • Automated calculation of statistics, application of confidentiality rules, and creation of quality indicators
  • Rigorous approved methods for standardization
• Adoption of GTAB has resulted in more standardization in disseminated products and practices within STATCAN
Future Developments

• Modernization initiative
• Open-data initiative
• Cloud-based storage
Thank you for attending!  
Merci de votre attention!

Questions?

Zixin Nie: zixin.nie@canada.ca  
Claude Girard: claude.girard@canada.ca