Cell Suppression as used to Protect Magnitude Data Tables

NCRN Virtual Seminar
April 1, 2015
Paul B. Massell
Mathematical Statistician
in CDAR; the
Center for Disclosure Avoidance Research
U.S. Census Bureau

Outline

- Background
- Why we needed to modernize the cell suppression software used in economic directorate
- How we met those challenges
- Survey and census tables that have been or will be protected with new software

Background 1

- Magnitude data tables are the main data product from the Economic Census and most economic surveys (conducted by Census Bur.)
- Typical magnitude variables are ‘sales’ and ‘number of employees’, ‘annual payroll’
- A typical cell value represents the sum of contributions from several establishments
- Each establishment is part of a firm (company)

Background 2

- If the contributions of 1 or 2 establishments comprise almost all of the cell value, cell is declared ‘sensitive’ and is suppressed
- Most tables are additive, so a single suppressed cell can easily be recovered
- Need to find additional cells, called ‘secondary suppressions’ to make exact recovery impossible

Background 3

- If a table has more than a few sensitive cells, secondary suppression software is required
- Bob Jewett was the main developer of a fine cell suppression program (in Fortran) that was used for over 20 years
- Gradually, a few weaknesses of the program were noticed. Economic Directorate decided it was time to modernize it

Challenge: Large Size of Tables encountered in Econ Directorate

- Many tables are large; some are huge
- Many tables have thousands of sensitive cells; some have millions
- Want to use LP (linear programming) model to achieve higher accuracy than network flow model achieves on 3D tables
- Problem: if not fine-tuned, program could take over 100,000 hours to run on a huge table
- To reduce time, must develop algorithms faster than those used in Jewett program

Challenges: Types of Tables encountered in Econ Dir

- DRB and Econ Dir decided many years ago, that C.B. must protect not just establishment values, but **company values** (i.e., sums of estab values associated with a company).
- This requirement is difficult to implement; it requires complex code. Some Econ divisions were not satisfied with the protection provided by Jewett program.

Challenges: Complex Geographies lead to many Linkages among Tables

- Each geographical relation of the form Geog Level 1 = sum of Geog Level2 (i) leads to a constraint that must be built into LP model.
- All overlapping relations must be processed on a single run.
- Jewett program used a method called ‘backtracking’ that was time-consuming; and had other bad aspects.

Meeting the Large Table Challenge 1

- ‘SKIP P’: While protecting a single sensitive (aka ‘P’) cell, notice if the associated ‘protection flow’ also provides protection for other P’s contained in suppression pattern

- ‘m at a time’: New algorithm uses a bit of parallelism; i.e., attempt is made to simultaneously protect a fixed number ‘m’ of P cells. If it leads to an infeasible solution, program protects this set sequentially.

Meeting the Company Level Protection Challenge

- Jewett program uses a complex notion called ‘capacity to protect’. It was only a partial solution. It failed to adequately measure the protection provided by a set of suppressed cells to each other.
- R&M group developed the notion of a ‘supercell’; a set of suppressions that lie in a ‘shaft’ in one dimension of a table. This provides good protection at company level.

Meeting the Challenge of Creating Useful Data Structures

- Program should be written in a language (e.g. C++) that allows for complex data
- One such data structure creates a graph into which all cell values and additive relations can be loaded. Nodes are used for cells; Arcs are used for relations.

- This structure makes it easy to identify disjoint sets of column relations, that we call ‘table groups’. A ‘table group’ represents a set of linked tables that are not linked to any others (in given table set).

Meeting the Large Table Challenge 2

- If a table group is very large, we may need to split it into partial ‘table groups’; and protect each partial group separately.
- Then pgm can be run on full table group with union of suppression patterns. Additional secondary suppressions may be needed.

Meeting other Challenges 1

- Handling Rounded Data
- Rounding of cell values usually leads to minor non-additivity of a table; LP model in supp pgm is not affected by this. This is because LP model requires only that **perturbations** of the values be additive.

Meeting other Challenges 2

- A nice feature of LP suppresion models, is that they are ‘self-auditing’ for protection patterns with standard assumptions. This means there is no need to check that required protection has been achieved.
- However, if table is unusual in some way (e.g., not additive) an enhanced audit program may be needed to check results.

Tables sets protected with new software

- Tables from ACES and BRDIS surveys, 2012 Econ Census Industry series, BasicL (geographic manufacturing series) has run and is under review, MECS (used an early version of pgm)
- Will be used for vast majority of Econ Census 2012 tables that undergo processing in next few months

References

- First paper describes the R&M group's work on cell suppression modernization
- U.N. Econ Conf ; Ottawa ...Oct. 2013
- FCSM Working Paper 22 on Disclosure Avoidance
- Massell..Company.Level.Protection..ICES.2012..paper
- Massell..Disclosure.Avoid.Agenda..FCSM.2013..slides
- Massell..Disclosure.Avoid.Agenda..FCSM.2013..paper