An Integer Programming Cell Suppression Algorithm – Providing Company Level Protection in One Optimization

Bei Wang
Economic Statistical Methods
Division, U.S. Census Bureau

Outline

- Cell suppression models and process
  - Linear Programming (LP) model
  - Sequential LP cell suppression program currently in production (LP-prod)
  - Integer Programming (IP) model
- Adding tailored capacity to the cell suppression model
  - LP-cap and IP-cap
- Applications of LP-cap and IP-cap

United States® Census Bureau

Background 1/2

Cell Suppression

- One of many disclosure avoidance methods
  - Primary sensitive cells are identified using p% rule - Ps
  - Complements are selected to protect the Ps from being derived via subtraction – Cs
  - Ps and Cs are suppressed in the publication
- Used by Economic Census and other economic programs

Background 2/2

Cell Suppression

- Model used
  - Network minimal cost flow (MCF) – 1992 through 2007
    - Protect one P and one hierarchical table at a time
    - Backtracking necessary for overlapping table, could lead to infinite loop.
  - Linear Programming (LP) – 2012, 2017, 2022
    - Adopted 2-trial – a suppression pattern is determined after two optimizations in each sequence
    - Solution is searched through whole data structure (globally)
    - Eliminated backtracking
    - Developed mLP – find protection for multiple primaries in one model
      - 1-LP – when m = 1

United States Census Bureau

Notation Used in the Model

Notation	Description
$x^{+}_{i,j,k}, x^{-}_{i,j,k}$	each cell defines two variables; one represents the flow coming into the cell and the other out of the cell
$x^{+}_{p,i,j,k}, x^{-}_{p,i,j,k}$	similar as above, but distinguished for each *P*
$z_{i,j,k}$	binary variables corresponding to $x^{\pm}_{p,i,j,k}$ used to coordinate across IP instances
i, j, k:	for a 3-dimension table with each index represents a dimension
$c_{i,j,k}$	constant, appears in cost objective function, determines the cost of cell
$cap^{p}_{i,j,k}$	capacity of cell*(i,j,k)* to target *p*
$v_{i,j,k}$	constant, usually cell's value, used in bound of a variable We sometime set $c_{i,j,k}$ = $v_{i,j,k}$

United States® Census Bureau

Other Notations

Notation	Description
P	set of primary cells determined by sensitivity rule, usually, p% rule
C	set of complements produced by cell suppression program
n_p	$= \|P\|$
Seq-LP	LP cell suppression program usually a sequential process
mLP	multiple (m) primaries are solved in a sequence
IP or sim-IP	IP cell suppression program referring a simultaneous process

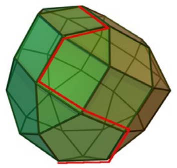
United States® Census Bureau

LP cell suppression model (used in LP-prod)

To find complementary (C)s for a particular primary (P)

- Objective

$$\min \sum_{i=1}^{rows} \sum_{j=1}^{cols} \sum_{k=1}^{levs} c_{i,j,k} \left(x^{+}_{i,j,k} + x^{-}_{i,j,k}\right)$$

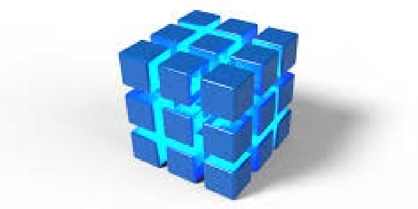


- Linear constraints
  - Additive_linear_constraints($x^{\pm}_{i,j,k}$):

$$\sum\nolimits_{rrel(i)} (x^{+}_{rrel(i),j,k} - x^{-}_{rrel(i),j,k}) = (x^{+}_{1,j,k} - x^{-}_{1,j,k}) \quad \triangleq \quad \text{for } j,k$$

$$\sum\nolimits_{crel(j)} (x^{+}_{i,crel(j),k} - x^{-}_{i,crel(j),k}) = (x^{+}_{i,1,k} - x^{-}_{i,1,k}) \quad \text{for } i,k$$

$$\sum\nolimits_{lrel(k)} (x^{+}_{i,j,lrel(k)} - x^{-}_{i,j,lrel(k)}) = (x^{+}_{i,j,1} - x^{-}_{i,j,1}) \quad \text{for } i,j$$



  - P_constraint:

$$x^{+}_{*} = prot(p)\,, \qquad x^{-}_{*} = 0$$

- Bounds $0 \leq x^{+}_{i,1,k},\ x^{-}_{i,1,k} \leq v_{i,j,k}$

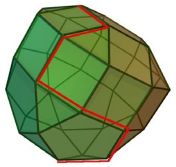
LP cell suppression program is a Sequential process (LP-prod, seq-LP)

- For each p in primary set {

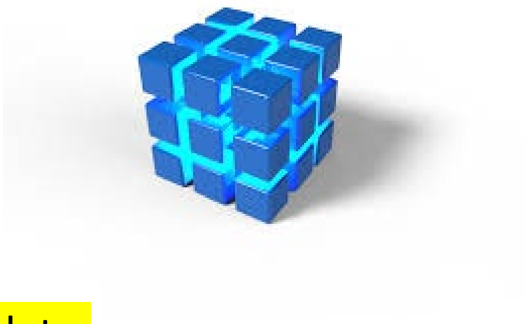
Solve

Is p protected?

objective:



subject to



+ p_constraints

Update:

1. Set Cs
2. Mark any additional Ps that are protected.

}

United States® Census Bureau

Simultanous IP model (Sim-IP)

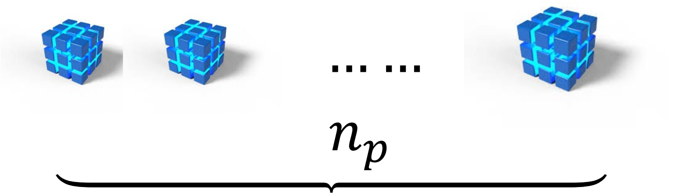
- Objective is to minimize value suppressed

$$\min Y = \sum_{i=1}^{rows} \sum_{j=1}^{cols} \sum_{k=1}^{levs} c_{i,j,k}\ z_{i,j,k}$$

- For each primary p_{index}

  additive_linear_constraints($x^{\pm}_{pindex,i,j,k}$),

  p_constraint(p_{index})



- For each primary p_{index}

  logic_constraints($z_{i,j,k}, x^{\pm}_{pindex,i,j,k}$):

$$x^{+}_{pindex,i,j,k} + x'^{-}_{pindex,i,j,k} - prot(p_{index}) * z_{i,j,k} \le 0$$

- Bounds($x^{\pm}_{pindex,i,j,k}$)

How is the company level protection achieved in LP-prod?

- Two passes:
  - Base pass provides table level protection
  - Super cell pass provides company level protection
    - Find list of super cell (sc)- aggregation of 2 or more of P&C cells, in a relation, that failed p% test.
    - For each sc
      - Check and potentially lower the capacity of cells in the dimension of supercell
      - Find complementary suppression(s)
- Why two passes? Why weren't the tailored capacity implemented in the LP-prod?
  - Allows one model to handle multiple Ps
  - Leads to over suppression

United States Census Bureau

Two-pass Programs

	One pass	Two-pass
LP Model (sequential)		LP-prod: mLP, 1LP
IP Model (simultaneous)		IP

United States® Census Bureau

Adding Tailored Capacity to Cell Suppression Model

- Capacity defines how much protection a cell can give to a target P cell
- Tailored capacity defines how much protection a cell can give to a target P cell *when accounting for company-level protection*
- Usually, capacity appears in the bounds

$$0 \leq x^{+}_{i,j,k},\ x^{-}_{i,j,k} \leq cap^{p}_{i,j,k}$$

$$0 \leq cap^{p}_{i,j,k} \leq v_{i,j,k}$$

- To 1-LP model

  Only one pass necessary, but n optimizations vs 2n from LP-prod

- To IP model

  Leads to a one optimization problem: one pass

Calculating Tailored Capacity

- We use *p*% rule to identify sensitive cell. When evaluating an adjacent cell, *x*, trying to protect a sensitive cell, we measure the aggregates by p% rule
  - Compute prot(*p*) and prot(*x* ∪ *p*) by *p*%
    - prot(*P*) – how much protection does the P need
    - prot(*x* ∪ *p*) – how much protection does *P* need if cell *x* is selected as complementary
  - If prot(*x* ∪ *p*)>0 then
    - $cap^p(x)$ = prot(*P*) – prot(*X* ∪ *P*)
  - Else
    - $cap^p(x)$ = val(*X*)

United States® Census Bureau

The one pass algorithm

- Implementing tailored capacity to the base pass
- Eliminating super cell pass
- Applies to both LP (1-LP) and IP (LP-cap & IP-cap)

Why one pass? What are the advantages?

1. more sufficient suppression
2. Ideal for IP cell suppression model

Programs

	One pass	Two-pass
LP Model (sequential)	LP-cap	LP-prod: mLP, 1LP
IP Model (simultaneous)	IP-cap	IP

↑ One optimization program

Example 2-pass vs 1-pass

Simple 1D data : column 1 = column 2 + column 3 + column 4 + column 5					
LP-prod (2-pass)					
1	2	3	4	5	
242	155 P=6	26 C=4 (super pass)	24 C=6 (base pass)	37	
LP-cap (1-pass)					
242	155 P=6	26 C=6	24	37	

United States® Census Bureau

IP Cell Suppression Model

- a perfect candidate to add tailored capacity

- A new set of variables is used for each $p \in \mathrm{P}$
- For example, $x^{\pm}_{p_1,i,j,k}$ for p_1, and $x^{\pm}_{p_2,i,j,k}$ for p_2 where $p_1, p_2 \in \mathrm{P}$
    - $x^{\pm}_{p_1,i,j,k}$ and $x^{\pm}_{p_2,i,j,k}$ referring to the same cell in the table for the same *I,j,k*, yet are two different variables.
  - The disjoint sets of variables allows us to add tailored capacity
- Adding tailored capacity to a simultaneous IP cell suppression model
  - Find solution only in one optimization
  - Provide desired company level protection

Test data

- Tiny

  A tiny extract from Econ Census - 70 cells & 42 Ps

- Annual Capital Expenditure Survey (ACES)

  4620 cells & 71 Ps

- Econ Census

  1958 cells & 891 Ps.

United States® Census Bureau

Tests on three different data sources comparing LP-prod and LP-cap		Tiny	2015 ACES	Econ Census
# Cells in the overall table or publication		70	4,620	1,958
# P's (primaries)		42	71	891*
LP (Standard for comparison)	Count of Complements	16	197	177
	Value of Complements	96,910	4,475,000k	3,425k
LP-cap	Count of Complements	14	190	177
	Value of Complements	94,650	4,471,000k	4,183k
%change in suppression (cells,value)	Count of Complements	-12%	-3%	-0%
	Value of Complements	-2%	-0.1%	+2%

*This is an unduplicated count
All data value truncated to 4 significant digits for disclosure purposes

Sources US Census Burean [Disclosure Review Board (DRB) approval number: CBDRB-FY24-ESMD005-002].]

Tests on three different data sources comparing LP-prod and IP-cap		Tiny	2015 ACES	Econ Census
# Cells in the overall table or publication		70	4,620	1,958
# P's (primaries)		42	71	891*
LP (Standard for comparison)	Count of Complements	16	197	177
	Value of Complements	96,910	4,475,000k	3,425k
IP-cap	Count of Complements	9	148	171
	Value of Complements	69,730	4,359,000k	3,367k
%change in suppression (cells,value)	Count of Complements	-44%	-25%	-3.4%
	Value of Complements	-28%	-2.6%	-1.7%
*This is an unduplicated count All data value truncated to 4 significant digits for disclosure purposes				

United States® Census Bureau

Sources US Census Burean [Disclosure Review Board (DRB) approval number: CBDRB-FY24-ESMD005-002].]

Conclusions

- Better suppression pattern, in general
  - LP-cap > LP-prod
- Although in one case
  - LP-cap < LP-prod (Econ Census – research in progress)
- IP-cap is a one optimization program that
  - Provides the best cell suppression results from all test data

    Examples are Annual Capital Expenditure Survey (ACES), Econ Census
  - Under suppression is possible when $cap^p(x_1 \cup x_2) < cap^p(x_1) + cap^p(x_2)$
  - Is computationally complex - NP-hard

Bibliography

- José H. Dulá, James T. Fagan, Paul B. Massell. (2004). *Tabular Statistical Disclosure Control:Optimization Techniques in Suppression and Controlled Tabular Adjustment.* Suitland: US Census Bureau https://www.test.census.gov/content/dam/Census/library/working-papers/2004/adrm/rrs2004-04.pdf
- Martin Serpell, Alistair Clark, Jim Smith and Andrea Staggemeier. (n.d.). Pre-processing Optimisation applied to the Classical Integer Programming Model for Statistical Disclosure Control.
- Philip Steel, James Fagan, Paul Massell, Richard Moore Jr., John Slanta, Bei Wang. (2013). Re-development of the Cell Suppression Methodology at the US. *Joint UNECE/Eurostat work session on statistical data confidentiality.* Ottawa, Canada.

I’d like to thank Godfried de Goey for his comments and suggestions.