*eCon*fidentiality

- a Disclosure Avoidance Application System (Proposed)

Bei Wang

U.S. Census Bureau

April 21, 2015

Outline

- Economic Census Disclosure Avoidance
  - Economic Census Background
  - Disclosure Avoidance Research
- Current Disclosure Avoidance Programs
  - Context
  - Methods
- Future - *eCon*fidentiality

Economic Census Background

- Tabulation goals
  - Industry-level summaries
  - Geographic Area Series
- Confidentiality concerns
  - We cannot "make any publication whereby the data furnished by any particular establishment or individual under this title can be identified"
- Magnitude of the problem
  - Large number of items
  - 18 Sectors
  - Approximately 1 million primary suppressions/per sector

Primary Cell Suppression: p-percent rule

- **Notation**

  T = cell total of absolute values of company data

  L = absolute value of data for largest company

  S =absolute values of data for second largest company

  rem = remainder = T - L – S

  p = p-percent value, e.g. p=25 for 25%

- Perform primary suppression if rem < L*p/100
  - *p* confidential value

Primary Suppression Example

Sales

Industry	County 1	2	3	
11	700	400	375	1475
22	1000	600	450	2050
33	100	375	650	1125
	1800	1375	1575	4650

Example: Industry 11, County 1

- For *p*-percent rule, let $p = 20$
- Applying p-percent rule
  - T = 375
  - Bob's Sales = L = 250 (largest company)
  - Joe's Sales = S = 100 (second largest)
  - rem = 375 – 250 – 100 = 25
  - L*p/100 = (20 × 250)/100 = 50
- Since L*p/100 = 50 > rem = 25, we cannot publish this tabulation cell
  - Additional protection needed = ceil(L*p/100 – rem) = 25

Complementary Suppression

Sales

Protection required = 25

Industry	County 1	2	3	
11	700	400	P	1500
22	1000	600	450	2050
33	100	375	650	1125
	1800	1375	1575	4650

2012 Economic Census Publications

- 1513 releases in total

- Disseminated on a flow basis
  - Advance report (National, 2 -3 digit NAICS)
  - Industry series (National, 2 -6 digit NAICS)
  - Geographic area series (Subnational, 2 – 7 digit NAICS)
  - Subjects/summary series (Differs by sector)
  - Zip codes (selected sectors)

- Challenging disclosure avoidance problem, as each new release could affect confidentiality of prior release(s)

History

1979- 1982	Census Bureau develops heuristic cell suppression methodology
1984	Census Bureau purchases Minimal Cost Flow (MCF) optimization software from University of Texas and begins exploring more rigorous cell suppression methods Ongoing cell suppression research with Linear Programming (LP) and Integer Programming (IP)
1990	Network (flow) program developed for 2 and 3 dimensional tables (known in-house as "Jewett programs")
2010	Additive noise proposed as an alternative to cell suppression ▪ Adopted for selected economic programs ▪ Not pursued for the Economic Census

Cell Suppression Modernization Project

2008	Established dedicated team ▪ Methodologists - to understand and explore alternative methods ▪ Programmers – to implement methods effectively
2010	Focus on documenting/understanding existing methods and transforming FORTRAN programs to C++
2011 +	Developed new Linear Programming (LP) methods

Disclosure-Avoidance Processing

- Primary suppression of a published cell total using the *p-percent rule* protects the largest company in the cell from calculations performed by the second-largest company.
  - Conducted before Cell Suppression Program (CSP)
  - Input program requires additional amount of "protection" (*protection required*)

- Secondary suppression of a published cell total prevents the use of the table's additive relationships to solve for primary-suppression cell totals – provides the *required protection*

Relationships

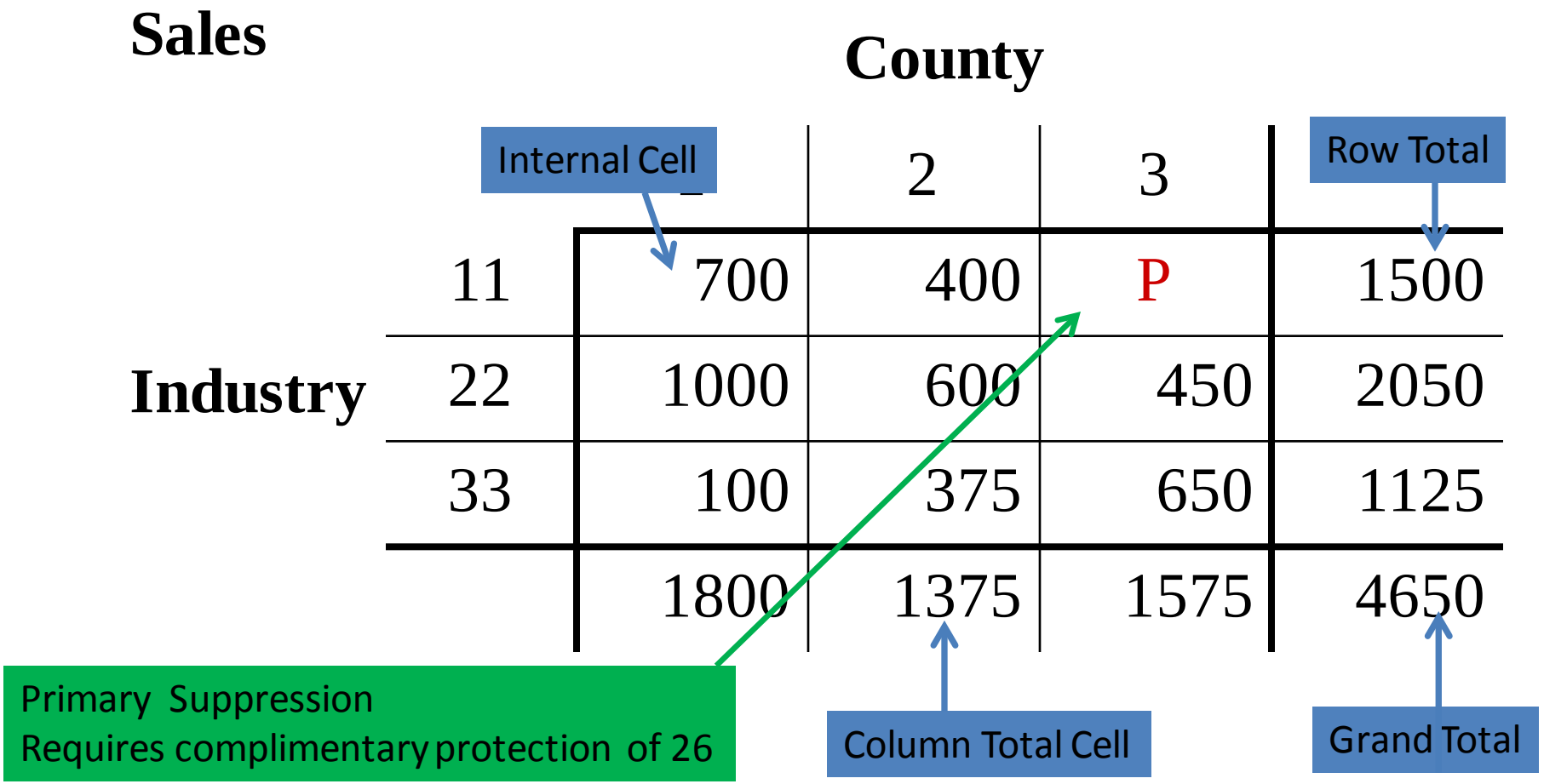
- Using “Sales” example
  - Row Relation R_{total} = 11 + 22 + 33
  - Column Relation C_{total} = 1 + 2 + 3

- Manufacturing geographic area series
  - 20658 separate geographic categories in 5222 relations
    example: state = $County_1$+$County_2$+…+$County_n$
  - 518 (up to 6-digit) NAICS in 154 relations
    example: 3113 = 31131+ 31134 + 31135
  - #tables 5222x154 = 804,188

Definitions and Concepts

- Cell
- Super Cell
- m-LP
- Skipping Ps
- Trials
- Parameters
  - α, β, costscale

Definitions and Concepts: Cell



Definitions and Concepts: Cell

- Cell Characteristics
  - Dimension
  - P/protection_required, C
  - cost/value
  - Capacity
  - Freeze
  - unpublished/publish
- Super Cell
  - Aggregate of cells; sensitive under aggregation

Definitions & Concepts: m-LP

- m-LP (Wang JSM 2013)
  - Protects m Ps simultaneously with one LP formulation
  - Adds m-1 additional pairs of constraints to model
  - m = 1 is the standard LP process

- A successful m-LP requires "well-grouped" m Ps that is feasible and achieves as much optimality as its 1-LP counterpart or better

- Time used is a fraction of 1-LP (=1/m of total 1-LP)

m-LP model

$$minimize: Y = \sum_{i=1}^{rows} \sum_{\substack{j=1 \\ (i,j,k)\in A}}^{cols} \sum_{k=1}^{levs} c_{i,j,k}\left(x_{i,j,k}^{(u)} + x_{i,j,k}^{(l)}\right)$$

subject to:

(a) $$\sum_{\substack{k=2 \\ (i,j,k)\in A}}^{levs}\left(x_{i,j,k}^{(u)} - x_{i,j,k}^{(l)}\right) = x_{i,j,1}^{(u)} - x_{i,j,1}^{(l)}$$

for i =1, ... , rows, j = 1, ... ,cols : $levs > 1$, $ws(i,j,1) = 0$

m pairs of (e)

(b) $$\sum_{\substack{i=1 \\ (i,j,k)\in A}}^{\lim r(ii)}\left(x_{rowrel(ii,i),j,k}^{(u)} - x_{rowrel(ii,i),j,k}^{(l)}\right) = x_{rowrel(ii,0),j,k}^{(u)} - x_{rowrel(ii,0),j,k}^{(l)}$$

for ii = 1, ... , rr, j = 1,..,cols, k = 1, ... , $levs$: $limr(ii) \geq 1$, $ws(ii,j,k) = 0$

(c) $$\sum_{\substack{j=1 \\ (i,j,k)\in A}}^{\lim c(jj)}\left(x_{i,colrel(jj,j),k}^{(u)} - x_{i,colrel(jj,j),k}^{(l)}\right) = x_{i,colrel(jj,0),k}^{(u)} - x_{i,colrel(jj,0),k}^{(l)}$$

for i = 1, ... , $rows$, jj = 1, ... , cc, k = 1, ... , $levs$: $limc(cc) \geq 1$, $ws(i,jj,k) = 0$

(d) $$0 \leq x_{i,j,k}^{(u)} \leq h_{i,j,k} \;;\quad 0 \leq x_{i,j,k}^{(l)} \leq h_{i,j,k}$$

for i = 1, ... , $rows$, j = 1, ... , col, k = 1, ... , $levs$: $(i,j,k) \in A$

(e) $$x_{prow,pcol,plev}^{(u)} = prot \;;\; x_{prow,pcol,plev}^{(l)} = 0$$

where:

$$c_{i,j,k} = \begin{cases} \max(0, v_{i,j,k}) & when\ (i,j,k) \in U \\ 0 & when (i,j,k) \in P \cup C \end{cases}$$

$$h_{i,j,k} = \max(0, v_{i,j,k})$$

Concepts: Skipping *P's*

- Often, providing additional protection to one targeted primary suppression (*P*) may protect additional P's
- Identifies the Ps that otherwise would result in a problem being done with objective=0 (no protection required) (*Steel et al 2013*)
- More than 99% of such primaries can be skipped. Time saved 99%, depending on the data

Concepts: Trials

A heuristic approach to optimize suppression between cells and value

- 1st trial establish a base pattern - for optimal value suppressed

- 2nd trial shake off the excesses by inverting the cost - for minimal number cells suppressed

- Example

c_1	c_2	c_3	total
10	60	100 (P = 20)	170

- Total Cost$(Trial_1)$ = $\begin{cases} 10 * 10 + 10 * 60 = 700 & if\ c_1, c_2 = C \\ 20 * 60 = 1200 & if\ c_2 = C \end{cases}$

- Total Cost $(Trial_2)$ = $\begin{cases} \frac{10}{10} + \frac{10}{60} = \frac{7}{6} & if\ c_1, c_2 = C \\ \frac{20}{60} = \frac{1}{3} & if\ c_2 = C \end{cases}$

1st trial choses c_1 & c_2 as complementary

2nd trial eliminates c_1

Parameters Controlling Cell Selection Behavior In Optimization

- ***Alpha*** (**α**) globally changes the relative cost of large and small cells - balancing between number and value suppressed (Wang JSM 2014), α Є (0,1]

- ***Beta*** (**β**) assigns flat cost to cells that are “freeze”, “unpublished”, “dummy”

- ***costscale*** assigns a proportional cost determined by end users data priority

Costscale Applied on Column Total

Sales

County

Important cells?

Cost double

Industry	1	2	3	
11	700	400	375	1475
22	1000	600	450	2050
33	100	375	650	1125
	1800	1375	1575	4650

α, β Applied on ASM2010

α applied

α	Total		Rating
1	2265	653445897	
.31	1705	649485679	++
.311	1760	6	
.312	1760	6	
.3125	1760	6	
.313	1760	6	
.314	1780	6	
.315	1794	6	
.35	1848	6	

505 cells saved from suppression

α β applied

α	Total		Rating	Published	
1	2005	687837136		665	574354598
.3125	1932	689723805	+-	558	572005027
.32	1906	686777987	++	560	570312702
.33	1885	684948227	++	557	571912231
.34	1899	689702800	+-	560	571935791
.35	1906	695174941	+-	560	572118398
.5	2028	712819567	- -	621	592225125
.8	2076	699797595	- -	697	581803417

Disclosure Avoidance Process

1. Gather requirement from subject area
2. Programmer runs cell suppression program
3. Subject area reviews suppression pattern
4. Revise requirements
5. Go to 2nd step (bottleneck)

Summary (Where We Are So Far)

Well-developed LP cell suppression methodology implemented in

- LP production software
- m-LP production software
- Used for 2012 econ census
- Advantages
  - Undersuppression eliminated
  - Oversuppression reduced
  - Speed almost satisfactory (need another 10x for big problems)
  - Program detects many data problems.
  - Program automatically decomposes data into the proper units for cell suppression.
  - User priorities can be addressed

Future enhancements/research

- R&M has a long list of items on the agenda
- A more robust m-LP

- Comparisons of Census LP system with others'
- Share with other agencies
- *eCon*fidentiality

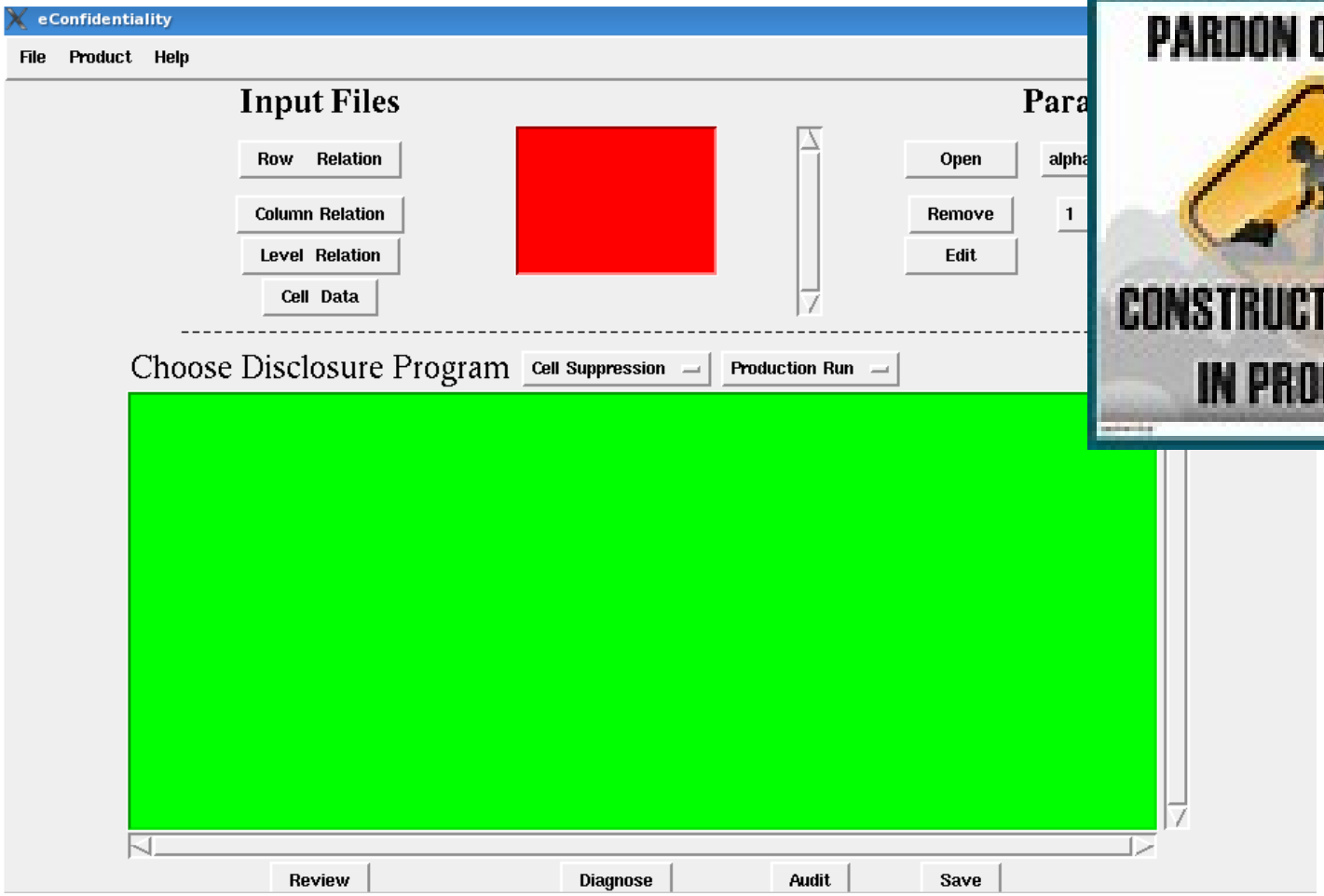
My GOAL: a User Controlled Process

- Current procedure
  - Users set up parameters
  - Programmers run programs
  - Users review output
    - Modify parameters as needed
    - Request additional program runs

- Vision – remove the “bottleneck” of programmers running the program

*eCon*fidentiality

Disclosure Avoidance Application System



References

- B. Wang Improve LP Process in Cell Suppression, Proceedings of the Government Statistics Section, American Statistical Association, Alexandria, VA (2013) CD-ROM
- B. Wang Using Weighting to Improve Cell Suppression Pattern in Annual Survey of Manufactures, Proceedings of the Government Statistics Section, American Statistical Association, Alexandria, VA (2014) CD-ROM
- P. Steel et al Re-development of the Cell Suppression Methodology at the US Census Bureau, UNECE Ottawa, Canada, 28-30 October 2013

Acknowledgements

Philip Steel, Katherine (Jenny) Thompson – presentation guidance

R & M team members: Paul Massell, Richard Moore, John Slanta, James Fagan, Phil Steel, Vitoon Harusadangkul – methodology and software development

Chris Chapman (BLS) - invitation and organization of session

Thanks!

Comments and Suggestions?

Bei.wang@census.gov