Exploring the Application of Differential Privacy to a Subset of Cells in a Table

Habtamu Benecha, Yang Cheng, Michael Jacobsen, John Grant, Lu Chen, Luca Sartore, Valbona Bejleri

National Agricultural Statistics Service

FCSM 2024

October 24, 2024

Disclaimer

The findings and conclusions of this presentation are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy.

Motivation

- NASS conducts the Census of Agriculture (CoA) every 5 years
  - Data published at national, state, and county levels
  - Network flow-based cell suppression system is used to protect census data
- Advancements in statistical disclosure limitation (SDL) research since the current NASS disclosure control approach was developed in 1990
- NASS is currently researching different SDL methods that use cutting-edge technologies
- One research direction focuses on exploring the application of noise-based methods to the CoA
  - Some of these methods apply noise to only a subset of cells of a table
    - Utility of data is preserved from unaltered cells
  - Transparency

Motivation

- Differential privacy (DP)
  - Transparent
  - Provides strong privacy protection
  - Several desirable properties
  - Utility can be affected because DP applies noise to all cells
- Some cells of a table may not require protection (i.e., non-sensitive cells) due to various reasons

Research Goal: Explore the feasibility of applying DP methods only to a subset of cells identified as "sensitive" in a table.

Identifying Sensitive Cells

- P-percent rule (FCSM Statistical Working Paper #22, 2005)
  - Cell suppression
  - Let U be the cell total, U_1 be the unweighted value for the largest respondent, and U_2 be the unweighted value for the second largest respondent.
  - The cell is sensitive if R < $U_1 \times P/100$, $R = U - U_1 - U_2$
  - P is determined by an agency
- Random Tabular Adjustment (RTA) (Stinner, 2018)
  - Based on Bayesian decision theory
  - Assumptions on the distributions
  - Utility is maximized while disclosure risk is bounded
    - Disclosure control parameter
    - Cells that require random noise are identified
    - Random noise generated from normal distribution

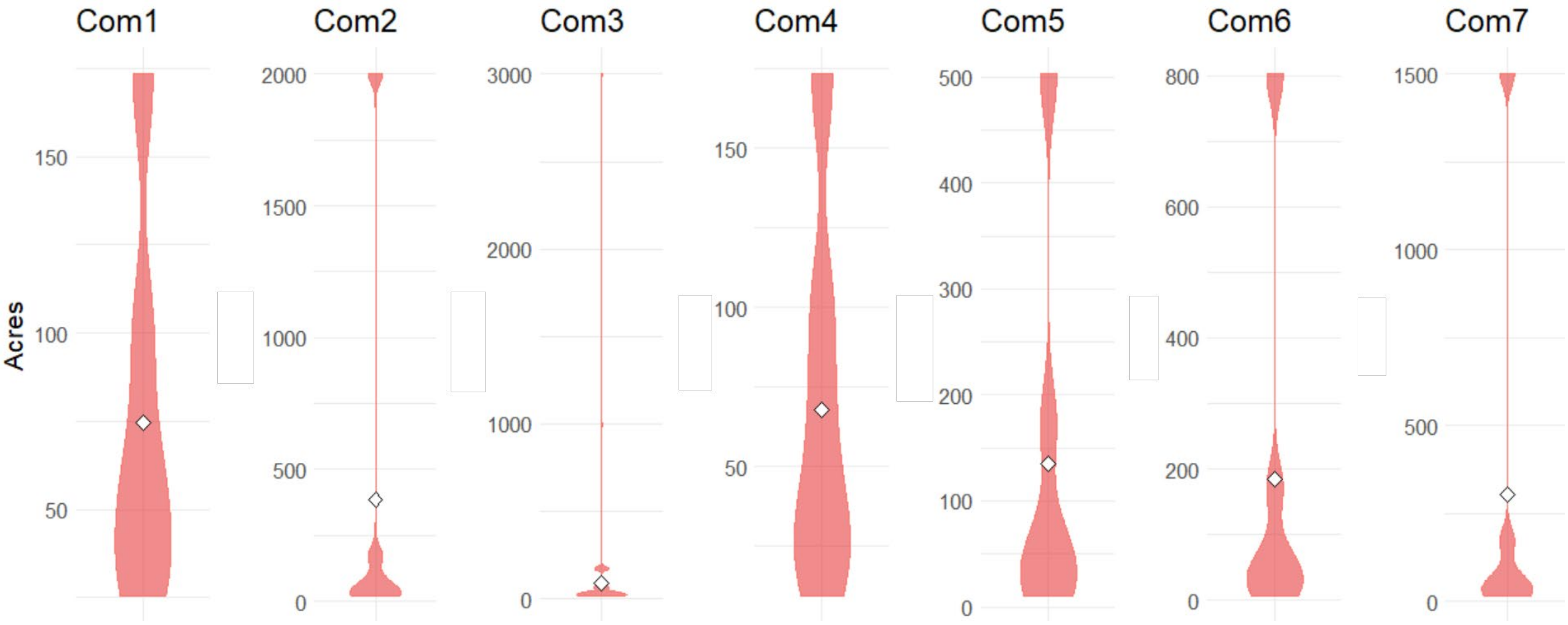
Differential Privacy & Per-record Differential Privacy (PRDP)

- Differential privacy
  - Privacy loss is bounded by the privacy budget ($\mathcal{E}$)
  - Aggregates (total sums) are often published
    - Sensitivity, Δf, can be very large
- A few farms can influence the amount of noise due to skewness in agricultural data
- To mitigates this problem: Per-record differential privacy (PRDP) (Seeman et al., 2023; Finley et al., 2024)
- PRDP: improved data utility with relaxed privacy guarantee to larger farms
  - Level of privacy guarantee varies from farm to farm
- Value of the privacy threshold, T, is dependent on the percentage of records that receive full DP guarantee

Example: Acreage Data by County and Commodity

- Harvested acres by commodity tabulated for counties & state
- Six counties, seven commodities
- An internal cell represents harvested acres of a commodity for a county
- Simulated microdata
- Respondent values for 41 of the 42 internal cells generated from a normal distribution with very small variances
  - Contributors to these cells have very close values
    - The p% rule may not identify these cells as sensitive
- 20 records per cell

Distribution of Commodity Acres in Microdata



Example: Acreage Data by County and Commodity

- One sensitive cell according to p% rule, p=20

Acres by county & commodity

County	Commodity							Total
	Com1	Com2	Com3	Com4	Com5	Com6	Com7	
A	2,006	40,001	**4,350**	1,995	9,998	15,974	30,000	104,325
B	1,000	391	420	209	241	158	299	2,718
C	3,438	3,450	3,441	3,446	3,444	3,441	3,442	24,102
D	1,298	1,298	1,298	1,305	1,303	1,294	1,299	9,094
E	665	658	662	658	656	666	660	4,625
F	535	537	535	539	546	543	537	3,772
State	8,942	46,335	10,706	8,151	16,188	22,076	36,237	148,636

An Application of Cell Suppression

- Four cells are suppressed when cell suppression is applied

Acres by county & commodity

County	Commodity							Total
	Com1	Com2	Com3	Com4	Com5	Com6	Com7	
A	2,006	40,001	D	1,995	9,998	15,974	D	104,325
B	1,000	391	420	209	241	158	299	2,718
C	3,438	3,450	3,441	3,446	3,444	3,441	3,442	24,102
D	1,298	1,298	1,298	1,305	1,303	1,294	1,299	9,094
E	665	658	662	658	656	666	660	4,625
F	535	537	D	539	546	543	D	3,772
State	8,942	46,335	10,706	8,151	16,188	22,076	36,237	148,636

An Application of RTA

- Only one internal cell needed random noise when RTA is applied
  - A total of four cells affected including 3 marginals
  - Assumptions for distributions

Acres by county & commodity

County	Commodity							Total
	Com1	Com2	Com3	Com4	Com5	Com6	Com7	
A	2,006	40,001	4,225	1,995	9,998	15,974	30,000	104,199
B	1,000	391	420	209	241	158	299	2,718
C	3,438	3,450	3,441	3,446	3,444	3,441	3,442	24,102
D	1,298	1,298	1,298	1,305	1,303	1,294	1,299	9,094
E	665	658	662	658	656	666	660	4,625
F	535	537	535	539	546	543	537	3,772
State	8,942	46,335	10,581	8,151	16,188	22,076	36,237	148,510

An application of PRDP

- All cells are altered
- $\epsilon = 2$; privacy threshold (T) was selected so that 50% of records receive full DP protection for each commodity

Acres by county & commodity

County	Commodity							Total
	Com1	Com2	Com3	Com4	Com5	Com6	Com7	
A	2,005	40,001	4,320	1,984	9,971	16,017	29,987	104,285
B	1,043	432	413	183	255	135	360	2,821
C	3,430	3,435	3,448	3,423	3,455	3,451	3,446	24,088
D	1,337	1,275	1,304	1,270	1,297	1,393	1,313	9,188
E	655	681	661	726	654	638	750	4,767
F	560	530	532	655	513	518	539	3,847
State	9,030	46,355	10,677	8,241	16,145	22,152	36,396	148,995

- DP mechanisms: higher noise values for sum queries on skewed data (Seeman et al., 2023)

Method Explored

- Given a dataset and an associated table to be protected
- Assume that some of the cells of the table are known to be non-sensitive (i.e., do not need protection)
- Proposed steps
  - Classify cells of the table in two categories based on sensitivity
  - Apply PRDP to the sensitive cells
  - Update the table by substituting only the sensitive cells with their altered values
  - Quality and Risk assessment
  - Publish the table
- Marginal totals of the table may change depending on noise added to sensitive cells

Case Study

- Table on sales of grains: 2017 CoA
- Counties/cells sum to the state total
- Grain categories: Corn, wheat, soybeans, sorghum, barley, other grains
- Only counties with at least three farms producing a grain are included in the analysis
- Table with 378 cells including marginal totals
- P% rule to identify sensitive cells (P=15)
- 33 primary & 19 secondary suppressions
- DP (Laplace noise), PRDP, and combination of P% rule & PRDP (P_PRDP) applied, $\epsilon = 2$
- PRDP: 50% of farms producing a commodity will receive full DP protection

Case Study

Number of cells in each category of absolute percent relative difference after noise is added

% Abs. Relative diff.	Number of cells		
	DP	PRDP	P_PRDP
0	0	0	345
(0, 5)	71	314	15
[5, 20)	95	45	12
[20 - 40)	44	8	4
[40 - 60)	15	4	1
[60 - 80)	18	2	1
[80 - 100)	8	0	0
>=100	127	5	0

$$\% \, Abs. Relative \text{ difference} = \frac{|Altered - Original| * 100}{Original}$$

Final Remarks

- Explored the application of combined SDL approaches to simple tables
- Utility sensitive to the method used for applying noise to the cell
- Level of privacy protection not studied
  - Overall, weaker privacy protection
  - Level of privacy from P_PRDP needs to be investigated

Future Work

- Assessment & quantification of disclosure risk
- Further research on sensitivity of cells
- Application to hierarchical & linked tables

Acknowledgment

We thank Ashwin Machanavajjhala & William Sexton from Tumult Labs for introducing us to the application of differential privacy & per-record differential privacy.

References

Cox, L. and Dandekar, R. (2002). Synthetic Tabular Data – An Alternative to Complementary Cell Suppression (unpublished manuscript).

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), 211-407.

Dulá, J.H., Fagan, J.T., Massell, P.B. (2004): Tabular Statistical Disclosure Control: Optimization Techniques in Suppression and Controlled Tabular Adjustment. Census Bureau Research Report. https://www.census.gov/content/dam/Census/library/working-papers/2004/adrm/rrs2004-04.pdf.

Evans, T., Zayatz, L., & Slanta, J. (1996). Using noise for disclosure limitation of establishment tabular data. In *Proceedings of the Annual Research Conference, US Bureau of the Census, Washington, DC* (Vol. 20233, No. 4, pp. 65-86).

Finley, B., Caruso, A. M., Doty, J. C., Machanavajjhala, A., Meyer, M. R., Pujol, D., ... & Terner, Z. (2024). Slowly Scaling Per-Record Differential Privacy. *arXiv preprint arXiv:2409.18118*.

Seeman, J., Sexton, W., Pujol, D., & Machanavajjhala, A. (2023). Privately Answering Queries on Skewed Data via Per Record Differential Privacy. *arXiv preprint arXiv:2310.12827*.

Stinner, M. (2018). Disclosure control and random tabular adjustment. *Disclosure*. Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference. https://nces.ed.gov/FCSM/pdf/G5_Stinner_2018FCSM.pdf.

FCSM (2005). Statistical Policy Working Paper 22, Second version; Report on Statistical Disclosure Limitation Methodology. Confidentiality and Data Access Committee 2005. https://www.hhs.gov/sites/default/files/spwp22.pdf

Thank You!

For questions: Habtamu.benecha@usda.gov