Using Differential Privacy to Advance Rural Economic Development

Applying Data Privacy and Confidentiality Methods to Industry Employment Data

Claire McKay Bowen, Ajjit Narayanan, and Corianne Payton Scally

January 28, 2021

Most economic data for rural communities are not publicly available because of privacy concerns surrounding the small counts of businesses and employees within certain industries. This means policymakers and researchers may lack the fundamental information needed to promote evidence-based economic development planning and investments for small rural economies. To address this problem, we apply modern data privacy methods to the Bureau of Labor Statistics Quarterly Census of Employment and Wages data to generate differentially private synthetic data by census tract. We alter the counts of business establishments and employees by industry. In the five nonmetropolitan counties we tested, these differentially private synthetic datasets reflect the true data reasonably well while maintaining privacy. This brief defines the problem in more detail, explains the approach to selecting counties and generating data, discusses findings, and proposes next steps to exploring data privacy improvements on small population data.

Rural Places and Data Privacy

The small population data problem is real. Responsible data owners have tried many methods to preserve the privacy of people represented in their data, including data suppression (not releasing data considered too easily identifiable) and publishing aggregate data at coarser levels, such as larger geographies. The data owners—such as federal statistical agencies—apply these aggressive data privacy approaches as more public data become available. Last year, the US Census Bureau attempted to link

records from the 2010 Census to other public data to identify individual records. Although the Census Bureau used several data privacy methods, its internal team discovered they could reidentify 45 percent of 2010 Census respondents using publicly available data sources such as Facebook.[1] At the same time, aggressive suppression and other coarse data privacy methods are less than ideal because they prevent or complicate accurate assessments of changes and impacts over time.

As a result of current data privacy methods, rural communities and stakeholders lack the fine-grained data they may need for economic development planning and investments (Scally, Burnstein and Gerken 2020). The inability to see how businesses and employment trends by industry change over time hinders future planning, such as how to encourage entrepreneurship and workforce development opportunities that fill gaps between what communities need and what services and jobs are available. These data could also help communities build strategically on local capacities, such as by identifying prevalent job skills that could align well with new business development activities.

Newer data privacy and confidentiality methods could improve the ability to release data for smaller geographic levels that reflect more accurate results while maintaining privacy. If we can maintain privacy or lower the disclosure risks using these methods, rural data users and practitioners could apply a wider range of statistical analyses or applications and better understand past and predict future business and employment trends.

Approach

The Bureau of Labor Statistics (BLS) Quarterly Census of Employment and Wages (QCEW) is a count of establishments,[2] employees, and wages reported by employers (BLS 2020). It contains 95 percent of US jobs by industry down to the county level. These data are generated primarily from state unemployment insurance programs and supplemented by other BLS surveys. The QCEW excludes proprietors, gig workers, unpaid family members, and some farm and domestic workers, among others. It also covers only some establishments on tribal lands.[3]

To see how differential privacy methods could allow us to make these data more usable for rural communities, we took the following approach:

- We chose a strategic sample of five nonmetropolitan counties to test our methodology.
- We generated a "ground truth" dataset of establishments and employees per industry at the census tract level. Because the QCEW only publishes these data at the county level, we imputed the tract-level data by fusing the QCEW data with external data sources.
- We synthesized new datasets by applying two algorithms: the Laplace sanitizer and the differentially private multinomial synthesizer.

Our approach resulted in three data privacy scenarios for testing:

1. **Baseline scenario:** The current BLS suppression disclosure method, which generates county-level counts of establishments and employees broken down by industry.

2. **Laplace sanitizer:** A common and easy non-parametric differentially private synthetic data method that generates census tract–level counts of establishments and employees broken down by industry.
3. **Differentially private multinomial synthesizer:** A common and easy parametric differentially private synthetic data method that generates census tract–level counts of establishments and employees broken down by industry.

County Selection

To select counties, we used the County Typology Codes published by the US Department of Agriculture's Economic Research Service (ERS) in 2015 and updated in 2017.[4] Of the 3,143 US counties and county-equivalents, 1,976 are categorized as nonmetropolitan, meaning they are outside the boundaries of a metro area (measured by population density and commuting patterns)[5] and have no cities with 50,000 or more residents.

The pool of eligible counties was reduced by the necessity to have counties with published QCEW data by industry and nonspecialized economies. To create our "ground-truth" dataset, we needed counties with QCEW data on establishments and employees by industry; we could not generate census tract–level data if county data were suppressed for many industries. And, to ensure that the data privacy results were broadly applicable across diverse economies, we selected from the 585 nonmetropolitan counties with nonspecialized economies, as designated by the ERS. This means their economies did not primarily depend on farming, mining, manufacturing, federal or state government, or recreation. These requirements eliminated many smaller nonmetropolitan counties, a notable limitation to our analysis that should be addressed in the future.

To select five counties from the 585 nonmetropolitan, nonspecialized counties in 2015, we applied the following criteria:

- **Population size:** Counties should represent a range of population sizes while having at least three census tracts to test: one larger nonmetropolitan county with more than 100,000 residents, one smaller county with 20,000–39,999 residents, and any remaining counties having 40,000–99,999 residents.
- **Regional diversity:** Counties should be representative across the US, from the Northwest and Southwest, to the Midwest, to the Northeast and Southeast.
- **Racial and ethnic diversity:** Counties should reflect the racial and ethnic diversity of rural populations, including at least one county with close to 50 percent of Black residents and one with close to 50 percent of Hispanic or Latino residents.
- **Additional attributes:** Counties should exhibit other important non-mutually exclusive demographic attributes important to policymakers, according to ERS County Typology Codes: at least one identified as a retirement destination, as having persistent poverty or persistent

child poverty, and/or as having residents with low education attainment; and at least one county with none of these additional attributes.

Finally, we excluded counties that contained tribal lands to avoid undercounting establishments. Table 1 summarizes the characteristics of the five selected counties.

TABLE 1

Characteristics of the Five Selected Rural Counties, 2015

County	Region	Population	Racial and ethnic diversity	Additional attributes
Ravalli County, MT	Northwest	41,902	93% white	Retirement destination
Chaves County, NM	Southwest	65,459	56% Hispanic/Latino	Low education, persistent child poverty
LaSalle County, IL	Midwest	110,401	85% white	None
Crisp County, GA	Southeast	22,846	44% Black	Low education, low employment, retirement destination, persistent poverty, persistent child poverty
Mercer County, WV	Northeast	60,486	90% white	None

Sources: 2015 USDA ERS County Typology and 2014–18 five-year American Community Survey data.
Notes: "Retirement destination" means the number of residents ages 60 and older increased 15 percent or more between 2000 and 2010. "Low education" means at least 20 percent or more of residents ages 25–64 lacked a high school diploma or its equivalent between 2008 and 2012. "Persistent child poverty" means 20 percent of more of related children under 18 years old were poor as measured by the 1980, 1990, and 2000 decennial censuses and 2007–11 five-year American Community Survey data. "Low employment" means less than 65 percent of residents ages 25–64 were employed in 2008–12. "Persistent poverty" means 20 percent or more of residents were poor as measured by the 1980, 1990, and 2000 decennial censuses and 2007–11 five-year American Community Survey data.

Generating "True" Tract Values

As noted above, the QCEW provides data on the number of establishments and employees at the county-industry level. For our project, we needed data at the census tract level, which are not publicly available but are held by BLS and state governments. We therefore imputed a "ground truth" dataset at the tract-industry level to allow for more granular analysis. We generated this tract-level ground truth data with the help of two external data sources: the proprietary InfoUSA 2017 business database and the 2017 Census Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES) data. We use 2017 data across all our data sources to make the comparison as equal as possible. The data fusion process involved three steps:

1. Using the external data sources to generate the proportions of establishment (InfoUSA) and employees (Census LODES) within each tract in a county
2. Multiplying the tract proportions by the total establishments and employees within that county reported by the QCEW

3. Rounding to non-negative integers and ensuring that the total establishments and employees within all tracts equal the county total

As an example, imagine a county with exactly three census tracts. Table 2 shows how we used the supplemental information from the InfoUSA data to generate tract-level establishment counts for the manufacturing industry, and table 3 shows how we used the supplemental information from the Census LODES data to generate tract-level employee counts for the manufacturing industry. In both tables, we used the QCEW county totals to impute the tract totals.

We then repeated these calculations for all tracts across all industries to generate a tract-industry–level dataset of the counts of establishments and employees within our five counties.

TABLE 2

Theoretical Manufacturing Establishment Counts, InfoUSA and QCEW

Tract	Number of establishments (InfoUSA)	Number of establishments (QCEW)	Number of establishments (imputed ground truth)
Tract 1	15	-	122 * (15/100) ~ 18
Tract 2	75	-	122 * (75/100) ~ 92
Tract 3	10		122 * (10/100) ~ 12
County total	**100**	**122**	**122**

TABLE 3

Theoretical Manufacturing Employee Counts, LODES and QCEW

Tract	Number of employees (LODES)	Number of employees (QCEW)	Number of employees (imputed ground truth)
Tract 1	200	-	1,200 * (200/1,150) ~ 209
Tract 2	250	-	1,200 * (250/1,150) ~ 261
Tract 3	700		1,200 * (700/1,150) ~ 730
County total	**1,150**	**1,200**	**1,200**

Applying Data Privacy and Confidentiality Methods

Once we had created our ground truth data, we tested three data privacy and confidentiality methods or statistical disclosure control methods: the BLS disclosure method used on the QCEW data, and two differentially private approaches that may generate more granular data while preserving privacy.

As of April 2020, BLS is updating its standards for statistical disclosure control methods, stating "the Bureau of Labor Statistics must formally protect establishment data that they collect by having *plausible deniability* for every value of sensitive data that they publish." In the context of the QCEW, plausible deniability means that if a sensitive record (e.g., number of employees) is altered for public release, then there should be enough uncertainty about the altered values that the associated employees could *plausibly deny* that the record refers to them. A popular formally private approach that provides plausible deniability is an algorithm that satisfies differential privacy (DP).

Dwork and others (2006) first proposed DP for quantifying the privacy-loss when releasing information from a confidential dataset. At a high level, DP links the potential for privacy-loss to how much the estimate for a unique statistic (or query) from the underlying confidential data changes with the absence or presence of any individual record that could be in the dataset. More specifically, DP quantifies the privacy-loss for each statistic by a parameter, ϵ, which is often referred to as the privacy budget. This parameter mathematically represents a bound on the log-odds for the probability that the protection scheme produces any particular output from a statistic given that any individual is in the data versus the probability that it produces the same output given that any individual is not in the data. In other words, privacy can be quantified (via the privacy budget), guaranteed, and composed across many analyses. This framework allows for a formal guarantee of the amount of information released about a confidential dataset over an arbitrary number of analyses, and it does not require assumptions concerning how a data intruder (someone who tries to extract sensitive information from publicly released data) would attack the data or the amount of information they possess.

Note that DP is a statement about the algorithm (or mechanism), not the data, which is a common misconception. In other words, DP is a definition and not a method, but a method can be differentially private.

BASELINE

The Bureau of Labor Statistics has three general disclosure limitation methods, which end up suppressing over 60 percent of the potential data cells at the county-industry level:

1. **Cell dominance:** Any cells that have less than three contributors are suppressed automatically.
2. ***P*-percent test:** This test determines how sensitive a cell is and whether it should be suppressed. We define the BLS *p*-percent test as follows:

   Let X be the target cell count, x_1 be the value of the largest contributor to X, c be the size of coalition (a group of respondents who pool their data to estimate the largest reported value), p be the pre-specified percentage ($0\% < p < 100\%$), and N be the total number of contributors to X. We can calculate the *p*-percent value by:

   $$S^{p\%}(X) = x_1 - \frac{100}{p}\sum_{i=c+2}^{N} x_i$$

   where $S^{p\%}(X)$ is the sensitivity measure of X for p percent. If $S^{p\%}(X) > 0$, then X is sensitive and should be suppressed. Note that c is typically 2, and if $N > 3$ then $S^{p\%}(X) > 0$ for any values of p.

   Because of disclosure risks, BLS does not report what value of p is used. This highlights a drawback to using *p*-percent: the lack of transparency.

3. **Secondary disclosure:** Secondary disclosure requires that any grouping of aggregate records at one level of the data cannot have only one cell suppressed. For instance, if a cell has four employers, but one was suppressed via the *p*-percent test, then it would be easy for an outsider

to calculate the number of observations from the fourth employer by going "up" one level. In this case, another cell or all cells in this level must be suppressed.

For some rural counties, applying the three methods suppresses the number of employees for specific county-industry combinations. Figure 1 shows the 12 county-industry combinations within our selected counties where the above rules caused the BLS to suppress the count of employees (the month_3_emplvl variable).

FIGURE 1

County-Industries Suppressed by BLS Disclosure-Limitation Methods, Ravalli County, Montana, and Crisp County, Georgia, 2017

	NAME	industry_title	disclosure_code	qtrly_estabs	month3_emplvl
	<chr>	<chr>	<chr>	<chr>	<chr>
1	Ravalli County, Montana	NAICS 22 Utilities	N	5	0
2	Ravalli County, Montana	NAICS 42 Wholesale trade	N	96	0
3	Crisp County, Georgia	NAICS 11 Agriculture, forestry, fishing and hunting	N	28	0
4	Crisp County, Georgia	NAICS 21 Mining, quarrying, and oil and gas extraction	N	1	0
5	Crisp County, Georgia	NAICS 22 Utilities	N	1	0
6	Crisp County, Georgia	NAICS 48-49 Transportation and warehousing	N	20	0
7	Crisp County, Georgia	NAICS 54 Professional and technical services	N	23	0
8	Crisp County, Georgia	NAICS 55 Management of companies and enterprises	N	1	0
9	Crisp County, Georgia	NAICS 61 Educational services	N	3	0
10	Crisp County, Georgia	NAICS 62 Health care and social assistance	N	64	0
11	Crisp County, Georgia	NAICS 71 Arts, entertainment, and recreation	N	1	0
12	Crisp County, Georgia	NAICS 72 Accommodation and food services	N	53	0

Source: Bureau of Labor Statistics (BLS) Quarterly Census of Employment and Wages, 2017.
Note: The ground truth data we use in this brief is at the tract-industry level, not the county-industry level shown here.

DIFFERENTIALLY PRIVATE SYNTHETIC DATA

Before describing the differentially private data synthesis methods we used for our simulations, we must describe the differentially private mechanism used to add noise to estimates in the data for us to generate the synthetic data. (We use "add noise," "sanitize," "alter," and "perturb" values interchangeably.) For a given value of ϵ, an algorithm or mechanism that satisfies DP will adjust the amount of noise added to the data based on the maximum possible change, given two databases that differ by one row or observation, of the statistic or data that someone wants to release. This value is commonly referred to as the global sensitivity (GS) (Dwork et al. 2006). The most common differentially private mechanism is the Laplace mechanism, which uses the GS to adjust the amount of noise added to the statistic (Dwork et al. 2006). Specifically, this mechanism adds noise to a statistic or query from a Laplace distribution, where the mean is 0 and the variance is the GS divided by ϵ. In other words, when GS is large or ϵ is small (less information released), more noise is added to the statistic. If the GS is small or ϵ is large (more information released), less noise is added.

Next, we describe the two proposed differentially private data synthesis methods and classify them into the same two categories used in Bowen and Liu (2020): non-parametric and parametric approaches. Non-parametric approaches are differentially private data synthesis methods that generate data from an empirical distribution; parametric approaches are algorithms that generate the synthetic data from a parameterized distribution or generative model. In other words, most non-parametric differentially private synthetic data techniques sanitize the cell counts or proportions from a

cross-tabulation of the data. Parametric differentially private synthetic data methods rely on estimating or learning an appropriate parameterized distribution based on the original data and sampling values from that distribution with noisy parameters.

Laplace sanitizer. The most basic non-parametric differentially private synthetic data method is the Laplace sanitizer (Abowd and Vilhuber 2008). For our data, this approach adds noise from the Laplace mechanism to all possible combination of counts on the number of establishments and employees. We define all possible combinations to include empty cells, which are a combination of attributes from our categorical variables and have no observations from the original/confidential data. We must incorporate these cells in our Laplace sanitizer method for our synthetic data to be considered differentially private.

We implement the Laplace sanitizer as follows:

1. Select the categorical variables under consideration for synthetic data generation (for our data, the categorical variables are the census tract and the industry code)
2. Identify all the possible categorical attribute classes
3. Calculate the number of observations within each attribute class
4. Add noise based on a given ϵ to the counts via the Laplace mechanism, where the GS is 1 for all counting queries

Differentially private multinomial synthesizer. Since our data are categorical, we could model the data based on a multinomial distribution and add noise or perturb the probability weights using the Laplace mechanism (Bowen and Liu 2020). We will refer to this method as a differentially private multinomial synthesizer, which is a simple parametric differentially private synthetic data method.

We apply the differentially private multinomial synthesizer as follows:

1. Select the categorical variables under consideration for synthetic data generation
2. Identify all the possible categorical attribute classes
3. Calculate the proportion of observations within each attribute class
4. Add noise to the proportion of observations via the Laplace mechanism, where the GS for proportion queries is $1/n$ such that n is the number of total observations
5. Draw the desired number of synthetic data observations using a multinomial distribution

CONSTRAINTS AND POST-PROCESSING

We applied a few constraints and post-processing steps to ensure the structure of the synthetic data was consistent. We verified these constraints with our BLS partners.

- When first generating the synthetic data, the number of establishments and employees summed to the same number as the original data *at the county level*. However, this can change with the next constraint.

- If the number of establishments equaled zero, we set the number of employees to zero. This is done to comply with the requirement that all employees in a tract must be associated with an establishment.

Assessing Utility of Results

We evaluated the utility or usefulness of the synthetic data by calculating summary statistics, using discriminant-based quality metric algorithms, and measuring the L1 distance between the synthetic and real data (i.e., how off are the counts?).

We compare the overall establishment and employee counts within each industry for each census tract and county between the original data (ground truth data) and the differentially private synthetic data. We provide this information visually as the difference in these counts from original data to the differentially private synthetic data.

GENERAL DISCRIMINANT–BASED QUALITY METRIC ALGORITHMS

We also applied a discriminant-based quality metric method, which is a technique that measures the overall distribution similarity between synthetic and original data. These metrics try to give a broad sense of how "close" the synthetic data are to the original data. More specifically, the approaches leverage propensity scores (predicted probabilities of group membership) to discriminate between the original and synthetic data, then use the estimated propensity scores to calculate the corresponding utility metrics in various ways. Researchers first developed these methods on traditional synthetic data, but they apply to differentially private synthetic data as well. At a high level, these utility measures train a classifier to discriminate between two datasets; the more poorly a classifier performs, the more similar distributionally the datasets are assumed to be.

We apply one discriminant-based quality metric algorithm, SPECKS (Bowen, Liu, and Su 2020). This approach requires training and fitting classifiers to the combined original and synthetic data with a binary indicator for whether the data row comes from the original or synthetic data. Each record in the original and synthetic data will receive a propensity score or a probability that it belongs to the synthetic data. We then determine the empirical cumulative distribution functions of the propensity scores from the original and synthetic data separately, and apply the Kolmogorov-Smirnov (KS) distance on the two functions. The KS distance is the maximum distance of the two empirical cumulative distribution functions, where the synthetic and the original data have the largest separation. A smaller KS distance (close to 0) indicates that the synthetic data preserved the original data well whereas a larger KS distance (close to 1) means the synthetic data differ a lot from the original data. For more on this method, please see Bowen, Liu, and Su (2018).

Preliminary Findings

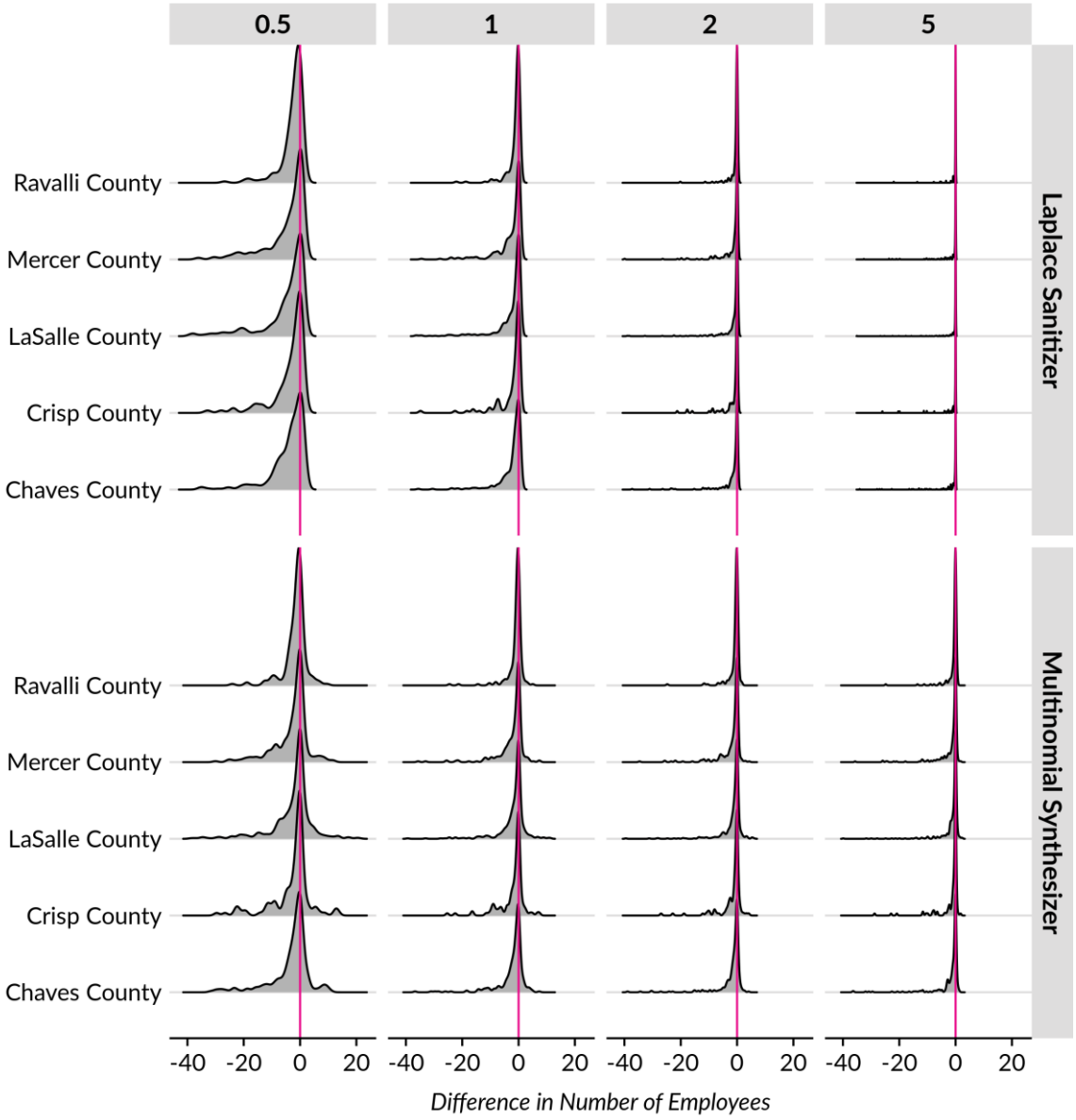
For our simulations, we set the privacy-loss budget as $\epsilon = \{0.5, 1, 2, 5\}$ and generated 1,000 differentially private synthetic datasets using both the Laplace sanitizer and the differentially private multinomial synthesizer. We conducted the repeats to gain a general sense of the average result.

Figures 2 and 3 show the difference in employment counts and establishment counts, respectively, as density distribution plots across all census tract–industry combinations. Overall, we see that the differences become smaller as ϵ increases for all methods and counties. The differentially private multinomial synthesizer seems to have more evenly distributed differences whereas the Laplace sanitizer is skewed to the left. This disparity in behavior likely stems from how the differentially private noise is added and how those values are constrained. For the Laplace sanitizer, noise is added directly to the counts; if a count is negative, then we adjust the value to 0. This procedure, called hard bounding, causes some bias in results (Bowen and Liu 2020). For the differentially private multinomial synthesizer, noise is added to the proportions and readjusted to ensure that the sum of the proportions is 1 and that all proportions are non-negative.

Figure 4 displays the overall number of employees and establishments and the KS distance from applying SPECKS. We applied two different models with SPECKS: a logistic regression or generalized linear model with all first-order interactions, and a Classification And Regression Tree model (commonly referred to as CART). We test two different models because each one varies in complexity and measures different types of distributional similarity. Similar to the results in figures 2 and 3, the utility improves as ϵ increases, except for the Laplace sanitizer where the KS distance increases at $\epsilon = 5$. One reason for this counterintuitive result is a combination of how noise is added to the counts and what constraints we place on the data. Another reason is the very small number of observations, which make it challenging for the two classifier models to perform proper predictions before trying to classify which records belong to the original and synthetic data. Overall, we see that the Laplace sanitizer performs better across most of our utility metrics for $\epsilon < 5$.

FIGURE 2

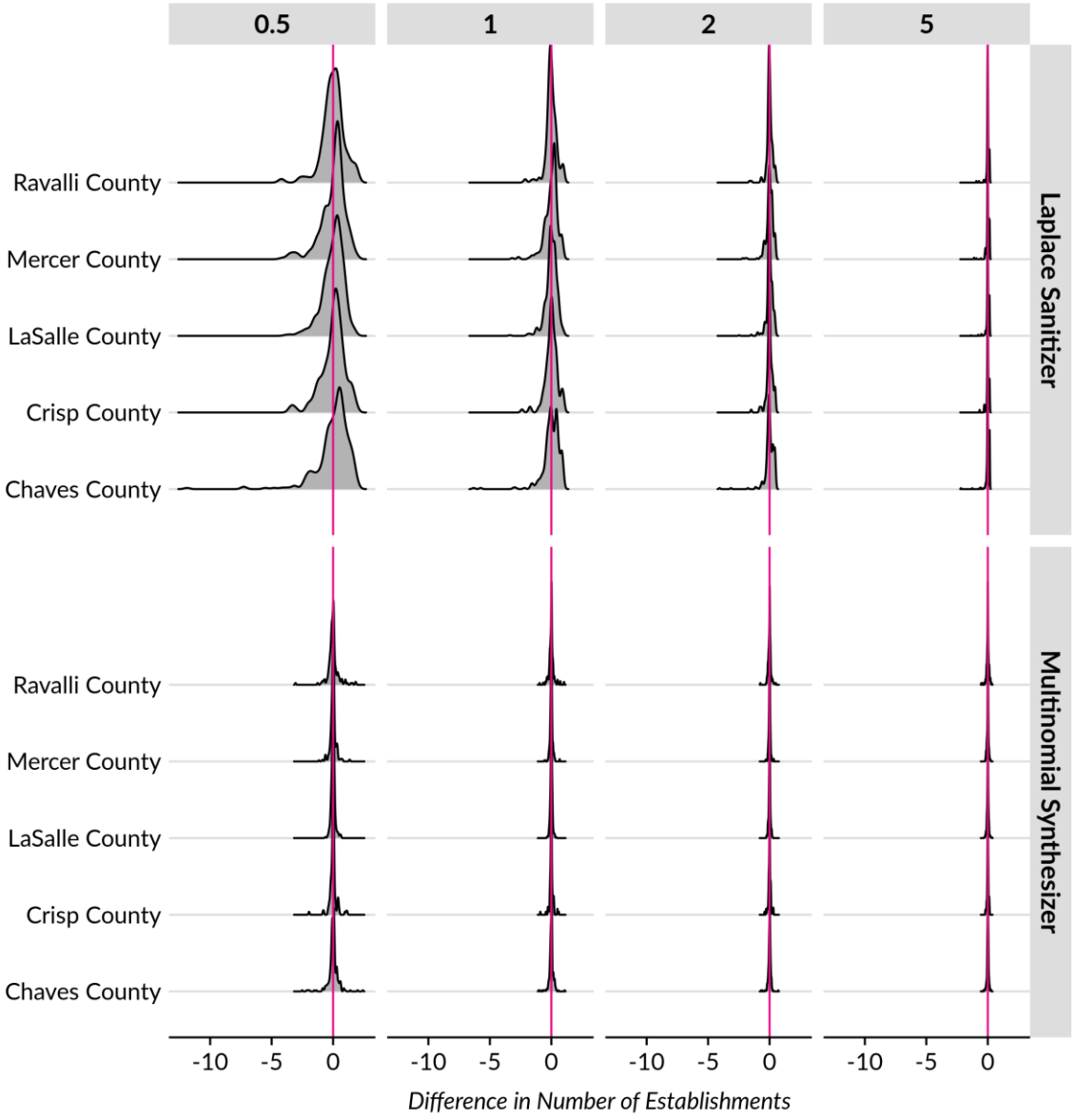
Difference in Tract-Level Employment Counts between the True and Synthetic Data for Five Rural US Counties



Note: This figure shows density distribution plots of the difference between the employee counts in the "ground truth" data and the employee counts in the synthetic data. These differences were calculated at the tract-industry level (i.e., manufacturing in tract 10) across 1,000 iterations and then turned into density distributions. Each row represents a county and differential privacy method, and each column represents the value of ϵ tested.

FIGURE 3

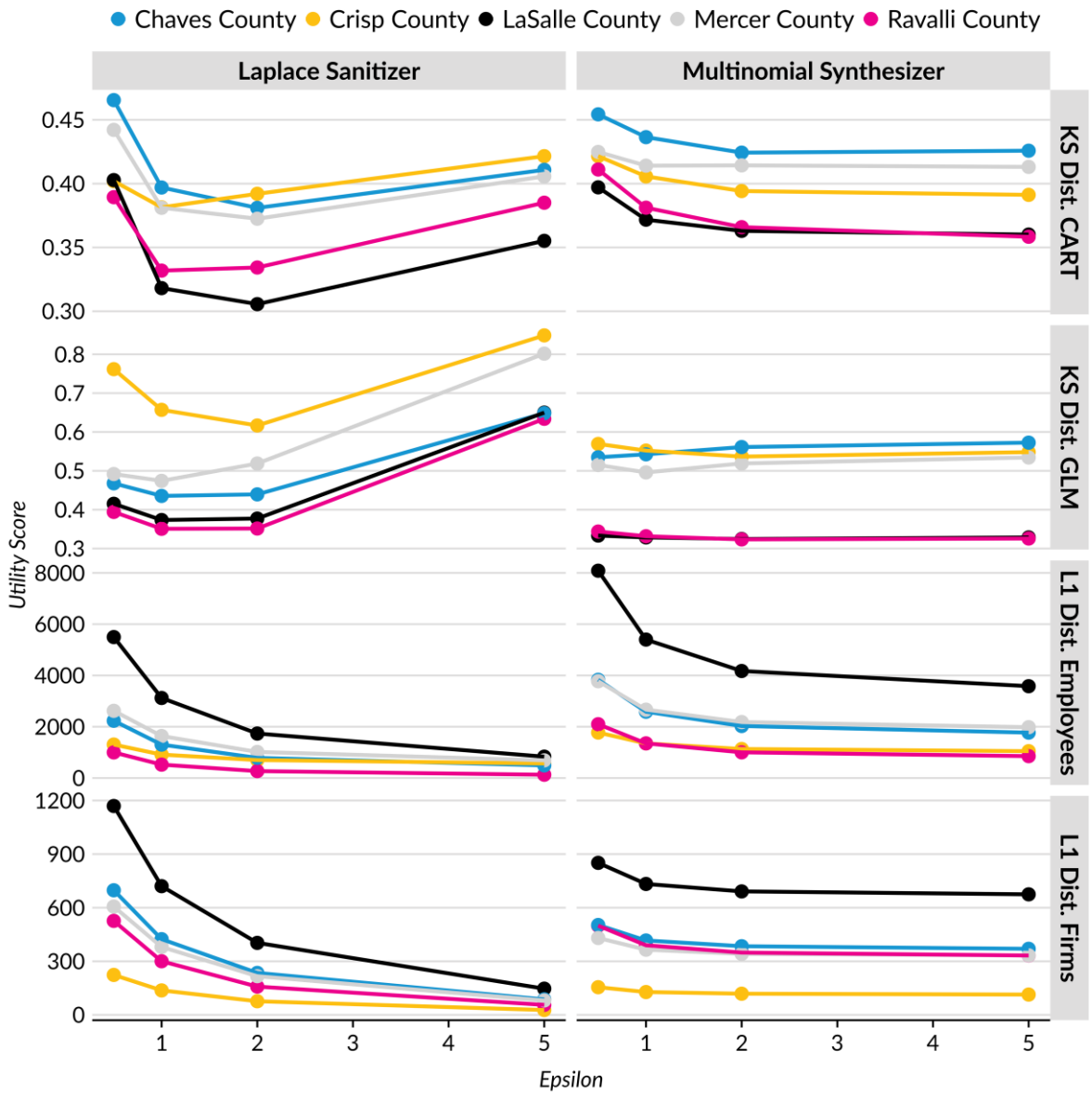
Difference in Tract-Level Establishment Counts between the True and Synthetic Data for Five Rural US Counties



Note: This figure shows density distribution plots of the difference between the establishment counts in the "ground truth" data and the establishment counts in the synthetic data. These differences were calculated at the tract-industry level (i.e., manufacturing in tract 10) across 1,000 iterations and then turned into density distributions. Each row represents a county and differential privacy method, and each column represents the value of ϵ tested.

FIGURE 4

Utility Metric Results for Five Rural US Counties



Note: This figure shows line plots of the various utility metric scores across all census tracts, industries, and counties. The x-axis represents the value of ϵ tested, each column is the differentially private synthetic data evaluated, and each row represents a utility metric measure.

Future Work

Although we accomplished our goal of applying differentially private data synthesis methods on rural economic data, we propose to further expand our research in the following ways:

1. We want to refine and develop a more sophisticated differentially private synthetic data method that can start from the state-level data and then go down to census tract–level or possibly microlevel data. This would also show how our methods work on larger, more populous counties.
2. We want to include the average wage variable in our data synthesis process in addition to employee and establishment counts. Wages data from the QCEW are often used by researchers but are also frequently suppressed by current BLS disclosure methods.
3. Another challenge we want to tackle is incorporating a seasonal component to the privacy framework. Seasonal employment changes occur often in rural establishments that rely on tourism, such as outdoor adventuring businesses. Since the QCEW is updated quarterly, our data synthesis methods need to preserve these trends across time.
4. We want to institute additional constraints on zero-count tracts. In talking with our BLS collaborators, we realized the importance of preserving the number of tract-industries with no firms or no employees. For example, an establishment going from two employees to zero indicates a business closure, which has special importance for researchers. We might consider matching the number of closures in future work.

During this project, we also contacted partners at the ERS to ask what they would do with the QCEW if it were made available at the census tract level. They stated that they would want to compare the differentially private synthetic data at the census tract level against the proprietary National Establishment Time-Series data, which contain establishments across all industries. They would be interested in using tract-level QCEW data immediately on some of their projects, including analyzing the impacts of broadband adoption and broadband programs on growth in numbers of business establishments and employment, and investigating relationships between the COVID-19 pandemic and business dynamics in rural areas.

Throughout this project, we have learned there is clear, strong appetite for more granular privacy preserving data, especially for data that tend to be sparse in rural communities. We hope that this case study is a starting point for work that democratizes rural economic data and preserves individual privacy.

Notes

[1] Seth Borenstein, "Potential Privacy Lapse Found in Americans' 2010 Census Data," Associated Press, February 16, 2019, https://apnews.com/article/aba8e57c145047b5bab11b62baaa7f7a.

[2] Establishments are a single economic unit, such as a mine, farm, factory, or store. Establishments are typically at one physical location and differ from a firm, or a company, which is a business and may consist of one or more establishments.

[3] Not all establishments owned and operated by Indian tribes or Alaska Native entities are required to file Unemployment Insurance tax and may not appear in administrative UI records. See agency guidance from the US Department of Labor here: https://oui.doleta.gov/dmstree/uipl/uipl2k1/uipl_1401.htm.

[4] "County Typology Codes," US Department of Agriculture Economic Research Service, accessed May 12, 2020, https://www.ers.usda.gov/data-products/county-typology-codes/

[5] Per the ERS, metro areas include all counties containing one or more urbanized areas (high-density urban areas containing 50,000 people or more). Metro areas also include outlying counties that are economically tied to the central counties, as measured by the share of workers commuting daily to the central counties.

References

Abowd, John M., and Lars Vilhuber. 2008. "How Protective Are Synthetic Data?" In *International Conference on Privacy in Statistical Databases*, 239–46. Berlin: Springer.

Bowen, Claire M., and Fang Liu. 2020. "Comparative Study of Differentially Private Data Synthesis Methods." *Statistical Science* 35 (2): 280–307.

Bowen, Claire M., Fang Liu, and Bingyue Su. 2018. "Differentially Private Data Release via Statistical Election to Partition Sequentially." *arXiv preprint arXiv:1803.06763*.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. "Calibrating Noise to Sensitivity in Private Data Analysis." In *Theory of Cryptography*, edited by Shai Halevi and Tal Rabin, 265–84. Berlin: Springer.

Scally, Corianne, Eric Burnstein, and Matthew Gerken. 2020. *In Search of "Good" Rural Data*. Washington, DC: Urban Institute.

About the Authors

Claire McKay Bowen is the lead data scientist for privacy and data security at the Urban Institute, where she develops methods for data privacy and confidentiality. Her research focuses on assessing the quality of differentially private data synthesis methods and creating better science communication.

Ajjit Narayanan is a data science analyst at the Urban Institute, where he works on research to advance equity and inclusion in rural and urban communities. He enjoys building open source data and mapping tools for researchers, government officials, and community advocates. He uses and advises on tools such as interactive mapping, big data platforms, and data visualization methods to relevant public policy issues.

Corianne Payton Scally is a principal research associate in the Metropolitan Housing and Communities Policy Center at the Urban Institute where she explores the design, implementation, and outcomes of affordable housing and community development policy and programs for vulnerable populations across US communities. Her research on rural communities ranges from measuring assets and capacity to guiding investments in equitable solutions that boost services and infrastructure, and improve health and economic mobility.

Acknowledgments

This brief was funded by the Urban Institute's 2020 Fleishman Innovation Award. The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

The authors thank Wendy Martinez and Daniell Toth from the United States Bureau of Labor Statistics and John Pender from the United States Department of Agriculture for taking the time to meet and chat with us, and for providing helpful comments throughout the entire project.

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is a leading research organization dedicated to developing evidence-based insights that improve people's lives and strengthen communities. For 50 years, Urban has been the trusted source for rigorous analysis of complex social and economic issues; strategic advice to policymakers, philanthropists, and practitioners; and new, promising ideas that expand opportunities for all. Our work inspires effective decisions that advance fairness and enhance the well-being of people and places.