Measuring the Degree of Difference in Perturbed Data

Marlow Lemons, Aref Dajani, Jiashen You, and John Jordan*

Abstract
Statistical agencies have an official responsibility to mitigate disclosure to protect respondent identity. Data swapping is a common technique to achieve that effort. Consequently, it is important to evaluate the quality of the perturbed data. We investigate several metrics to quantify the degree of discrepancy between two tabulated data sets. This list ranges from established statistics such as the Gini index to Shannon entropy and more heuristic metric like the effective swap rate. A simulation study compared distributions of these statistics under different settings of swap rate and skewness. Applications to the one-year American Community Survey are presented.

Key Words: Disclosure avoidance, data swapping, chi-square, heterogeneity comparison, categorical data analysis, American Community Survey

1. Introduction

In addition to fulfilling federal mandates to protect respondent confidentiality, federal agencies are also interested in measuring the quality of the data that they release. Record linkage and reidentification studies are prime and common procedures that measure data utility and assess disclosure risk. However, one common way of measuring data utility is evaluating the amount of perturbation between the released and original data. Data swapping, noise infusion (Abowd et al., 2012), topcoding (Duncan et al., 2011), and cell suppression (Kelly et al., 1992) are common disclosure avoidance methods to protect respondent identity, but consequently produce a resulting dataset with loss of information. It is of interest to researchers to quantify this loss of information and determine which measures would best quantify the loss.

Past studies have investigated different methodologies for measuring information loss in large-scale data. Domingo-Ferrer and Torra (2001) addressed how the mean square error, mean absolute error, and mean variation, which utilize covariance and correlation matrices, can measure information loss in continuous variables. These authors also presented and demonstrated several metrics for categorical variables, including the distance function, the entropy-based information loss measure, and the alternative information loss measure that are versatile for several disclosure avoidance methods. Within the U.S. Census Bureau, other methods have been applied to measure information loss and data utility. Steel & Zayatz (2001) used coverage estimation methods to compare tables from swapped verses unswapped data involving race for several geographies in the 2000 Census. In three evaluation studies by Lemons et al. (2013), Lemons & Freiman (2013a), and Lemons & Freiman (2013b), the researchers used effective swap rate, chi-square, and an index of homogeneity to identify states with the most perturbation in race, age, and Hispanic origin variables due to data swapping from the 2010 Census. These metrics helped the researchers rank the states and territories that were least and most affected by swapping observations at several geographic levels.

* *United States Census Bureau, Center for Disclosure Avoidance Research, Washington, D.C. 20233.* This presentation and the paper are released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

This study focuses on addressing the relationship among several metrics that measure data utility after perturbation involving categorical variables, namely the chi-square statistic (CS), effective swap rate (ESR), deviance statistic (DS), Gini difference index (GD), Shannon difference index (HD), and Rényi difference index (HD_3). There are two research questions surrounding this study. First, which metrics are related when measuring data perturbation? Then, how does the underlying distribution of a variable of interest (VOI), its number of levels, and the amount of perturbation affect the relationship of these metrics? Simulations for various conditions of the number of tracts, population sizes within tracts, and frequencies for the variable of interest help address these questions. We conclude with applying these metrics to data from the 2011 American Community Survey (ACS).

2. Methodology

The simulation study consisted of three stages: (1) *generating the data*, (2) *performing the perturbation*, and (3) *computing the metrics*. These processes are described in the proceeding subsections.

2.1 Generating the Data

A program using SAS 9.4 simulated 1,000,000 records to represent households living within a state. The households were clustered into ten counties evenly and then further into tracts, with tract sizes ranging from 5,000 households to 25,000 households depending on the number of tracts per county. From there, household tabular counts were computed at tract and county levels for a variable of interest.

2.2 Performing the Perturbation

After simulating the data, a percentage of records were randomly chosen for perturbation. The perturbation method for this study was data swapping. In this process, two records are selected at a time and a subset of these chosen pairs are interchanged. For more details on the data swapping process, refer to Navarro et al. (1988) and Ramanayake & Appelbaum (2010). It is important to note that records were randomly swapped, with the requirement that records swapped outside of their tracts, but not necessarily out of their counties.

2.3 Computing the Metrics

Table 1 shows a list of the six metrics computed to measure the data perturbation. Let f_{ij} and $f_{ij}^{'}$ be frequencies associated with category i and level j before and after perturbation respectively, with r_{ij} and $r_{ij}^{'}$ representing relative frequencies.

These metrics were chosen since they assume increasingly greater values perturbed data deviates from its original data and are considered popular statistics for measuring data utility across several fields of study (Giancristofaro & Bonnini, 2007). One of the more familiar metrics is the CS, which compares the frequencies between perturbed and unperturbed data. The ESR is the smallest proportion of records that could have been swapped from the unperturbed data (Lemons et al., 2013). Similar to the CS, small deviations in the frequencies between the two datasets would result in smaller values for the ESR. The DS is a common metric for measuring goodness of fit and carries similar properties to the CS statistic. Small frequencies from the unperturbed data can result in inflated values for the DS. The GD derives from the commonly used Gini index of heterogeneity, G, used in applied economics studies (Gini, 1912; Lerman & Yitzhaki, 1984) for measuring "equidistribution" in categorical data, and is defined as

Table 1: Metric Formulas

Metrics	Formula
Chi-Square	$\text{CS} = \sum_i \sum_j (f'_{ij} - f_{ij})^2 / f_{ij}$
Effective Swap Rate	$\text{ESR} = \frac{1}{2} \sum_i \sum_j \lvert f'_{ij} - f_{ij} \rvert / \sum_i \sum_j f_{ij}$
Deviance Statistic	$\text{DS} = 2 \sum_i \sum_j (f'_{ij} \log(f'_{ij}/f_{ij}))$
Gini Difference Index	$\text{GD} = \sum_i \sum_j [(r'_{ij})^2 - (r_{ij})^2]$
Shannon Difference Index	$\text{HD} = \sum_i \sum_j (r'_{ij} \log r'_{ij} - r_{ij} \log r_{ij})$
Rényi Difference Index	$\text{HD}_3 = \frac{1}{2} [\log \sum_i \sum_j (r'_{ij})^3 - \log \sum_i \sum_j (r_{ij})^3]$

$$G = 1 - \sum_i \sum_j r_{ij}^2. \tag{1}$$

The HD is a difference of the Shannon entropy index (Shannon, 1948), H, defined as

$$H = -\sum_i \sum_j r_{ij} \log r_{ij}, \tag{2}$$

and its application has been used in information theory (Giancristofaro & Bonnini, 2007). The HD_3 comes from the generalized index of entropy proposed by Rényi (1966) and defined as

$$H_\alpha = \frac{1}{1-\alpha} \log \sum_i \sum_j r_{ij}^\alpha, \tag{3}$$

where $\alpha = 3$. Note that the GD, HD, and HD_3 metrics represented differences in the G, H, and H_3 values between the perturbed and unperturbed data.

2.4 Plan of Analysis

The processes described in Sections 2.1 through 2.3 were replicated 1,000 times for each combination of five factors.

The first factor was the number of tracts within each county (*5*, *8*, and *10*). The second factor was the distribution of the number of households within a tract (*skewed* and *uniform*). The third factor was the number of levels for the VOI (*2* and *5*). These were considered since popular variables like gender and race-ethnicity typically contain this many levels. The fourth factor was the distribution of the number of households into the levels of the VOI (*uniform* and *skewed*). The fifth factor was the amount of perturbation classified as *Low*, *Medium*, and *High* for confidential reasons. The perturbation method used in this study was data swapping since it is a common disclosure avoidance method used within the Census Bureau (Lauger et al., 2014).

To answer the first research question, Spearman pairwise correlation coefficients were calculated among the six metrics from the simulations. Metrics with significant correlations (absolute value of 0.7 and higher) averaged across all factors were grouped, and these groupings were then used to report other Spearman correlations for more detailed cases.

To answer the second question, a five-way Analysis of Variance (ANOVA) main-effects model was created for each metric to identify significant factors at the 5% level of significance. A stepwise regression method, using a model entry probability of 0.10 and a model removal probability of 0.05, will be used to simplify the model. Diagnostic statistics, like the coefficient of determination (R^2) and the root mean square error (RMSE), were reported as measures to quantify model performance.

2.5 Application of Study

A second component of this study involved an application to real data, which consisted of two parts. First, the six metrics from Table 1 were computed on the head of householder age and head of householder race variables at the tract level for the state of Pennsylvania on 2011 ACS household data. The purpose is to compare these results to the results from the simulation. Metric values at the state level were then computed to understand how an increased number of levels affects these metric values. Age contained five levels, namely *16-24 years*, *25-34 years*, *35-44 years*, *45-64 years*, and *65 or older years*. These levels were chosen not only to match the number of levels from the simulation study, but to identify age cutoff values that are of interest.

In the second part, the six metrics from Table 1 were then computed on the head of householder age, race/ethnicity, and gender variables, as well as their three-way combination at the state level for the states of North Dakota, Kentucky, and Illinois in the 2011 ACS data. These states were chosen since they represent the 5th, 50th, and 95th percentiles with respect to state population. The race/ethnicity variable was divided into five levels, namely *White (non-Hispanic)*, *Black (non-Hispanic)*, *Other Race (non-Hispanic)*, *Multi-Race (non-Hispanic)*, and *Hispanic*. *Male* and *female* were the only levels for the gender variable. The age variable used the same levels aforementioned.

3. Results

3.1 Simulation Results

Table 2 contains the mean, standard deviation, and cutoff values for the 5th and 95th percentiles (trimmed ranges) of the six metrics stratified by the number of levels for the VOI. Several patterns were found from these statistics. The means for each metric increased as the number of levels increased with the exception of the GD (from 0.00087 to 0.00077). The amount of variability in the CS and DS metrics increased as the number of levels for the VOI increased. Variability decreased for the GD, HD, and HD_3 metrics, with the GD showing the largest decrease (0.00724 to 0.00260) and the HD_3 showing the smallest decrease (from 0.000166 to 0.000154). Despite the variability decrease for these metrics, overdispersion was observed since the standard deviation was higher than the mean. Trimmed ranges for the CS and DS metrics were not only similar, but increased as the number of levels increased. However, the trimmed ranges for the GD, HD, and HD_3 showed a right shift as the number of levels increased.

Table 3 contains the average Spearman correlation value across all factors. Spearman correlations ranged between 0.41 (between the ESR and GD metrics) to 1.00 (between the CS and DS metrics). Correlations of 0.7 and higher were considered to be strong relationships between the two metrics. For example, the CS had the strongest relationships with the DS (0.99) and ESR (0.92) metrics, while the weakest relationship were with the HD (0.58), GD (0.43), and HD_3 (0.42) metrics respectively. Results from this table suggest that the CS, ESR, and DS metrics have exhibit similar patterns in their values and should be grouped, while the GD, HD, and HD_3 metric represent a second group with similar

Table 2: Descriptive Statistics of Metrics

Metric	Levels	Metric Statistics		
		Mean	Std. Deviation	Trimmed Range
CS	2	14.84	10.73	(2.88 , 35.49)
	5	39.12	28.20	(7.91 , 92.47)
ESR	2	$z.$	$z.$	$z.$
	5	$z.$	$z.$	$z.$
DS	2	14.84	10.73	(2.88 , 35.53)
	5	39.17	28.23	(7.91 , 92.48)
GD	2	0.00087	0.00724	(-0.00915 , 0.01217)
	5	0.00077	0.00260	(-0.00268 , 0.00498)
HD	2	0.00110	0.00838	(-0.01011 , 0.01454)
	5	0.00258	0.00460	(-0.00614 , 0.01575)
HD_3	2	0.000025	0.000166	(-0.000174 , 0.000235)
	5	0.000054	0.000154	(-0.000136 , 0.000272)

[†]Results protected under Title 13 of the U.S. Code.

behavior. These findings were expected as the first three metrics are based on frequencies while the other three metrics are based on relative frequencies.

Table 3: Overall Spearman Metric Correlations Averaged Across All Factors

Metrics	Metrics				
	ESR	DS	GD	HD	HD_3
CS	0.92	0.99	0.43	0.58	0.42
ESR		0.92	0.41	0.54	0.43
DS			0.43	0.58	0.42
GD				0.91	0.90
HD					0.85

Table 4 contains the ranges of Spearman correlations observed from metrics within Group A (denoted in the AA column), metrics within Group B (the BB column), and metrics across the two groups (the AB column). These ranges were stratified by the level of perturbation, the tract distribution, and the distribution of the levels for the VOI. From the results in Table 3, the CS, ESR, and DS metrics were classified in Group A and the GD, HD, and HD_3 metrics were classified into Group B. Correlations between the metrics in Group A were highly positive with ranges being very small across all three factors. Correlations remained positive for metrics in Group B, but the ranges varied. The strongest Spearman correlations were observed when the tract and level distributions were skewed (at least 0.93 for low perturbation, 0.94 for medium perturbation, and 0.93 for high perturbation). These

conditions were also where the shortest ranges were observed. The smallest correlations, along with the largest ranges, occurred when the tract distribution was skewed and the level distribution was uniform.

Table 4: Spearman Correlations between Metrics by Perturbation

Perturbation	Tract Dist.	Level Dist.	Metric Groupings		
			AA	AB	BB
Low	Uniform	Uniform	(0.99 , 1.00)	(0.53 , 1.00)	(0.53 , 0.90)
	Uniform	Skewed	(0.99 , 1.00)	(0.51 , 0.99)	(0.59 , 0.94)
	Skewed	Uniform	(0.95 , 1.00)	(0.37 , 0.98)	(0.31 , 0.87)
	Skewed	Skewed	(0.95 , 0.99)	($<$ -0.01 , 0.06)	(0.93 , 0.96)
Medium	Uniform	Uniform	(0.99 , 1.00)	(0.54 , 1.00)	(0.54 , 0.90)
	Uniform	Skewed	(0.99 , 1.00)	(0.52 , 0.99)	(0.60 , 0.93)
	Skewed	Uniform	(0.95 , 1.00)	(0.37 , 0.98)	(0.31 , 0.88)
	Skewed	Skewed	(0.95 , 0.99)	($<$ -0.01 , 0.11)	(0.94 , 0.95)
High	Uniform	Uniform	(0.99 , 1.00)	(0.53 , 1.00)	(0.53 , 0.91)
	Uniform	Skewed	(0.99 , 1.00)	(0.51 , 0.99)	(0.60 , 0.94)
	Skewed	Uniform	(0.95 , 1.00)	(0.39 , 0.98)	(0.32 , 0.88)
	Skewed	Skewed	(0.95 , 0.99)	(-0.02 , 0.13)	(0.93 , 0.95)

A = {CS, ESR, DS}, B = {GD, HD, HD_3}

Unique patterns in correlations and ranges occurred when comparing metrics across the two groups. Although correlations remained positive when at least one of the distributions was uniform, the ranges were were slightly larger than the ranges observed from the metrics in Group B. Uniquely, the smallest correlations from these ranges came from comparing the GD to all three metrics in Group A. Additionally, negative to weak positive correlations (along with smaller ranges) were discovered when the tract and level distributions were skewed. The GD metric produced negative correlations with the ESR during low and medium perturbations (-0.0038 and -0.0012 respectively), and with all three metrics for high perturbation (-0.0075 with the CS, -0.0200 with the ESR, and -0.0038 with the DS metric).

3.2 ANOVA Model Estimates

Model results from the five-way ANOVA for each metric are outlined in Table 5. The baseline scenario is: "High" for perturbation, 10 tracts per county, 5 levels for the VOI, "Uniform" for the tract distribution, and "Uniform" for the level distribution. Bolded estimates denote significant estimates at the five percent significance level. Values for the R^2 ranged between 0.0145 for the GD model to 0.9327 for the ESR model. The lowest observed R^2 values were from the GD (0.0145), HD (0.0462), and HD_3 models (0.0549).

For the CS model, low and medium levels of perturbation had a significant negative effect on the CS compared to the high level. Significant negative effects also occurred when five or eight tracts per county were used compared to ten tracts per county, suggesting that less tracts significantly decreases CS values. The number of levels for the VOI had a significant negative effect on the CS value compared to five levels. Finally, skewed tract

Table 5: Main Effects ANOVA Model Results

Factors	CS		ESR		DS		GD		HD		HD_3	
	β	SE	β	SE	β	SE	β	SE	β	SE	β	SE
Intercept (α)	51.67		$z.$		51.68		0.00108		0.00300		0.000073	
Perturbation												
Low	**-35.85**	0.11	**-0.00110**	0.0000015	**-35.87**	0.11	**-0.00108**	0.000049	**-0.00245**	0.000072	**-0.000053**	0.0000014
Medium	**-17.84**	0.11	**-0.00045**	0.0000015	**-17.85**	0.11	**-0.00056**	0.000049	**-0.00124**	0.000072	**-0.000027**	0.0000014
Tracts												
Five	**-13.05**	0.11	**-0.00050**	0.0000015	**-13.05**	0.11	**-0.00038**	0.000049	**-0.00099**	0.000072	**0.000008**	0.0000014
Eight	**-8.13**	0.11	**-0.00018**	0.0000015	**-8.14**	0.11	**-0.00045**	0.000049	**-0.00104**	0.000072	**-0.000011**	0.0000014
Num. Levels												
Two	**-24.31**	0.09	**-0.00066**	0.0000012	**-24.32**	0.09	**0.00011**	0.000040	**-0.00148**	0.000059	**-0.000029**	0.0000012
Tract Dist.												
Skewed	**11.49**	0.09	$<$ -0.00001	0.0000012	**11.50**	0.09	**0.00089**	0.000040	**0.00194**	0.000059	**0.000044**	0.0000012
Level Dist.												
Skewed	**13.39**	0.09	$<$ -0.00001	0.0000012	**13.41**	0.09	**0.00012**	0.000040	**0.00103**	0.000059	**-0.000026**	0.0000012
R^2	0.7774		0.9327		0.7771		0.0145		0.0462		0.0549	
RMSE	11.587		0.00016		11.602		0.005		0.008		0.000156	

or level distributions significantly increases CS values. None of the factors were removed after applying stepwise regression.

Some results from the ESR model were similar with those from the CS model. Perturbation continued also had a significant negative effect on ESR values, with low perturbation resulting in smaller ESR values than high perturbation. Also, the number of tracts and the number of levels factors had significant negative effects on ESR. Tract and level distribution factors did not have a significant effect on ESR. These factors were also removed after applying the stepwise selection method. However, the values of the RMSE and R^2 remained unchanged (0.00016 and 0.9327 respectively).

Levels for all five factors were significantly different from their baselines in the DS model. Estimates for the perturbation, number of tracts, tract distribution, and level distribution followed the same patterns as those from the CS model. No factors were removed from the model following the stepwise regression method.

Although the factors from the GD model were all significant, some patterns were different compared to the previously described model. Perturbation had negative effects with low perturbation yielding significantly lower GD values than the high perturbation case. The number of tracts per county had a significantly negative effect on GD values. Counties containing five or eight tracts tend to have GD values that are 0.38 $\times 10^3$ and 0.45 $\times 10^3$ units less than counties containing ten tracts. This is different from the behaviors observed from the CS, ESR, and DS models. Another interesting discovery was that smaller levels for the VOI resulted in significantly higher GD values. According to the model, GD values were 0.11 $\times 10^3$ units higher when the VOI had two levels versus five. Similarly to the behavior of the CS and DS models, the tract and level distribution factors were statistically significant. No factors were removed after stepwise regression was applied.

In the HD model, each factor was statistically significant and none of the factors were removed after stepwise regression. Perturbation tends to decrease HD values by 2.45 $\times 10^3$ and 1.24 $\times 10^3$ units compared to high perturbation. Counties containing five and eight tracts significantly decreased HD values by 0.99 $\times 10^3$ and 1.04 $\times 10^3$ units compared to counties containing ten tracts. HD values were approximately 1.48 $\times 10^3$ units lower for cases where the VOI had two levels compared to five. Finally, GD values from data with skewed tract and level distributions were 1.94 $\times 10^3$ and 1.03 $\times 10^3$ units higher than their respective uniform baselines.

From the HD_3 model, perturbation had a negative effect on HD_3 values. These estimates suggest that more perturbation tends to decrease the values for this metric compared to high perturbation. However, data with counties having five tracts each have HD_3 values that were 0.01 $\times 10^3$ higher than data with counties containing ten tracts each. However, data with counties containing five tracts each have HD_3 values that were 0.01 $\times 10^3$ lower than data with counties containing ten tracts each. The number of levels factor had a significantly negative effect on HD_3 values when the two-level group is compared to the five-level group. Finally, the tract distribution factor had a positive effect on HD_3 values when comparing skewed data to uniform data ($\beta = 0.04$), while the level distribution factor had a significantly negative effect on HD_3 when comparing the skewed level to the uniform level. The stepwise regression method did not remove any of the factors.

3.3 Application Results: Pennsylvania

Table 6 contains tract-level metric results analyzed on gender and age variables for the state of Pennsylvania in the 2011 ACS. The data for this state consisted of 67 counties and 3,185 tracts containing at least one housing unit within them. It is important to note one pivotal reason why the means and percentile values from all six metrics were well below

those observed from Table 2. The data swapping procedure used in the simulations of this study involved random swapping, but the data swapping procedure used for the ACS data performs swaps based on target variables that identifies at risk records in the data (see Lauger et al. (2014)).

Table 6: Age & Race Metric Results for Pennsylvania

VOI	Metric	Metric Groupings		
		Mean	Std. Deviation	Trimmed Range
Gender	CS	0.09	0.22	(0.00 , 0.43)
	ESR†	*z.*	*z.*	*z.*
	DS	0.08	0.24	(0.00 , 0.40)
	GD	-0.00027	0.02680	(-0.02000 , 0.02078)
	HD	-0.00027	0.02364	(-0.02043 , 0.02153)
	HD_3	-0.000584	0.043495	(-0.049777 , 0.048819)
Age	CS	0.26	0.51	(0.00 , 1.17)
	ESR†	*z.*	*z.*	*z.*
	DS	0.21	0.67	(0.00 , 1.23)
	GD	-0.00010	0.02439	(-0.02721 , 0.02497)
	HD	-0.00197	0.03011	(-0.05260 , 0.05099)
	HD_3	-0.000787	0.056656	(-0.087008 , 0.083861)

†Results protected under Title 13 of the U.S. Code.

Table 7 provides metric values for age, race/ethnicity, gender, and their three-way combination at the state level for Pennsylvania in the one-year 2011 ACS. The CS estimates ranges from 544.14 for gender alone to 8,017.87 for the three-way combination. DS values ranged from 541.29 for gender to 6,135.06 for the three-way combination. The results from the CS and DS metrics suggested that their values increase as the number of cells created increased. Although the statistics for the ESR are protected under federal law, the results followed a similar pattern to those of the CS and DS metrics. The GD metric values ranged from -1.19 (gender) to 16.44 (race/ethnicity). Ironically, the value of the GD for the three-way combination was smaller than that of race/ethnicity alone. The HD values ranged between -4.14 and 25.83 over the three demographic variables and their interaction. The HD also experienced the case in which the result for the three-way interaction (-4.14) was smaller than age, race/ethnicity, and gender results alone. The HD_3 showed similar patterns with the CS metric, with HD_3 values ranging between -0.0010 (for gender) to 0.0071 (for age$\times$race/ethnicity$\times$gender).

3.4 Application Results: State Level Results

Table 8 contains calculations of the six metrics on age at the state level for North Dakota, Kentucky, and Illinois. Values for the CS ranged between 61.83 and 930.41, 63.23 and 890.51 for the DS, 0.15 and 2.57 for the G, 0.47 and 5.53 for HD, and 0.0023 and 0.0031 for the HD_3. These values were computed based on 53 counties and 205 tracts for North Dakota, 120 counties and 1,115 tracts for Kentucky, and 102 counties and 3.123 tracts for Illinois. Results from this table suggest that states with higher populations experience higher metric values.

Table 7: Demographic Variable Comparisons for Pennsylvania

Metrics	Age	Race/Ethnicity	Gender	Age × Race/Ethnicity × Gender
CS	1,080.57	1,593.79	544.14	8,017.87
ESR[†]	$z.$	$z.$	$z.$	$z.$
DS	1,027.24	1,617.98	541.29	6,135.06
GD	2.03	16.44	-1.19	3.44
HD	4.47	25.83	-1.29	-4.14
HD_3	0.0025	0.0047	-0.0010	0.0071

[†]Results protected under Title 13 of the U.S. Code.

Table 8: State Level Metric Comparisons for the Age Variable

Metrics	North Dakota	Kentucky	Illinois
CS	61.83	303.62	930.41
ESR[†]	$z.$	$z.$	$z.$
DS	63.23	295.29	890.509
GD	0.15	0.72	2.57
HD	0.47	1.44	5.53
HD_3	0.0023	0.0023	0.0031

[†]Results protected under Title 13 of the U.S. Code.

4. Conclusions

The purpose of the study was to understand the relationship between several metrics that measure data utility after perturbation by determining those metrics that were highly correlated and determining the significant factors that explain each metric. Results from this study indicates that the CS, ESR, and DS metrics measure data quality and loss of information similarly, while the GD, HD, and HD_3 metrics measure similarly. The CS, ESR, and DS metrics correlate well regardless of the amount of perturbation, and also correlate well when either the tract or level distributions are uniform. For skewed conditions, the GD, HD, and HD_3 metrics correlate well.

Results also suggest that the amount of perturbation applied to the data significantly affects all six metrics in the same way. The more that data are perturbed, the higher the metric value. With exception to the GD and HD metrics, decreased number of tracts results in decreases in the metric value. VOIs with a smaller number of levels tends to have smaller metric values. Finally, skewed distributions within the data yield higher values in five of the six metrics.

There are several areas for future work. Domingo-Ferrer and Torra (2001) described several perturbation methods for disclosure avoidance, including multiple imputation, data synthesizing, rank swapping, rounding, and the post-randomization method. It would be of interest to readers to investigate how such methods would affect the correlations between the six metrics. Another is to compare the metrics from this study on geographies with very different land areas but similar populations. An example is the city of Washington versus the state of Wyoming. The results in this study noted that low chi-square scores indicated

a higher level of data usability. Is there a maximum threshold for each metric in which stakeholders are satisfied with? This would indicate the fulcrum where effective disclosure avoidance does not come at the expense of data usability.

5. Acknowledgments

The authors would like to thank Amy Lauger from the Center for Disclosure Avoidance Research, as well as Mark Asiala from the Decennial Statistical Studies Division of the United States Census Bureau, for their support to make this research possible.

References

Abowd, J. M., Gittings, R. K., McKinney, K., Stephens, B., Vilhuber, L., & Woodcock, S. D. (2012). Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time series. *U.S. Census Bureau Center for Economic Studies Paper No. CES-WP-12-13*.

Domingo-Ferrer, J. and Torra, V. (2001). Disclosure control methods and information loss for microdata. In *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies* (eds. Doyle P., Lane J. I., Theeuwes J.J.M. and Zayatz L.), pp. 91-110: North-Holland, Amsterdam.

Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R., & Roehrig, S. F. (2001). Disclosure limitation methods and information loss for tabular data. *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, 135-166.

Duncan, G. T., Elliot, M., & Salazar-González, J. J. (2011). *Why Statistical Confidentiality*? Springer: New York, NY.

Giancristofaro, R. A., & Bonnini, S. (2007). Permutation tests for heterogeneity comparisons in presence of categorical variables with application to university evaluation, *Metodoloski Zvezki*, *4*(1), 21.

Gini, C. (1912). Variabilitá e mutabilitá. Reprinted in *Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T)*. Rome: Libreria Eredi Virgilio Veschi, 1.

Kelly, J. P., Golden, B. L., & Assad, A. A. (1992). Cell suppression: disclosure protection for sensitive tabular data. *Networks*, *22*(4), 397-417.

Lauger, A., Wisniewski, B. & McKenna, L. (2014). Disclosure avoidance techniques at the U.S. Census Bureau: Current practices and research, *Technical Report 2014-02*, Center for Disclosure Avoidance Research U.S. Census Bureau.

Lemons, M., Freiman, M. & Dods, J. (2013). An evaluation of disclosure avoidance for the 2010 decennial census with respect to the age variable, In the 2013 *Census Confidential Report Series: Disclosure Avoidance #2013-03*.[1]

Lemons, M. & Freiman, M. (2013a). An evaluation of disclosure avoidance for the 2010 decennial census with respect to the race variable, In the 2013 *Census Confidential Report Series: Disclosure Avoidance #2013-03*.[1]

[1]Confidential papers protected under Title 13 of the U.S. Code (circulated on a need to know basis only).

Lemons, M. & Freiman, M. (2013b). An evaluation of disclosure avoidance for the 2010 decennial census with respect to the Hispanic origin variable, In the 2013 *Census Confidential Report Series: Disclosure Avoidance #2013-03*.[1]

Lerman, R. I., & Yitzhaki, S. (1984). A note on the calculation and interpretation of the Gini index. *Economics Letters*, *15*(3), 363-368.

Navarro, A., Flores-Baez, L., & Thompson, J. (1988), Results of data switching simulation. In Spring meeting of the *American Statistical Association and Population Statistics Census Advisory Committees*.

Ramanayake, A. & Appelbaum, S. (2010). An evaluation of the ACS swapping procedure: 2005-2008, In the 2010 *Census Confidential Report Series: Disclosure Avoidance #2010-01*.[1]

Reiss, S. P. (1982). Data swapping: a technique for disclosure control. *Journal of Statistical Planning and Inference*, *6*(1), 73-85.

Reiss, S. P. (1984). Practical data swapping: the first steps. *ACM Transactions on Database Systems*, *9*(1), 20-37.

Rényi, A. (1966). *Calculus des probabilités*, Dunod, Paris.

Shannon, C. E. (1948). A mathematical theory of communication, *Bell System Technical Journal*,*27*, 623-656.

Steel, P. & Zayatz, L. (2001). How variations in geography and changes in race coding affect disclosure prevention. In the 2001 *Census Confidential Report Series: Statistical Research Division #2001-01*.[2]