NISS

Distortion Measures for Categorical Data Swapping

Shanti Gomatam and Alan F. Karr

Technical Report Number 131
January, 2003

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Distortion Measures for Categorical Data Swapping

Shanti Gomatam[*] and Alan F. Karr[†]
National Institute of Statistical Sciences
Research Triangle Park, NC 27709–4006
{sgomatam,karr}@niss.org

Abstract

Data swapping is a common technique for statistical disclosure limitation, but its effects on real data are not understood completely. In this paper, we consider measures that can be used to quantify distortion to the data engendered by data swapping when the variables in the data set are categorical. These measures are applied to a data set derived from the Current Population Survey. Their behavior is studied and compared for various values of the swapping rate and different choice of the variable swapped.
Key words: data utility; data confidentiality; statistical disclosure limitation; Hellinger distance; Shannon entropy; total variation distance; contingency coefficient; Cramer's V.

1 Introduction

With the increase in information collected and maintained, especially by various federal agencies, and the increased demand for access to such information, the use of statistical disclosure limitation methods to protect confidentiality is essential and widespread. In the usual scenario, data are collected on subjects (individuals or organizations) and then transferred to the disseminator/agency, under conditions of protecting confidentiality. The disseminator then makes the data available to users, making sure that the data are protected in such a

[*] S. Gomatam is also Assistant Professor, Department of Mathematics, University of South Florida, Tampa, FL.

[†] The authors thank Ashish Sanil and Adrian Dobra for their comments and suggestions. Support for this research was provided by National Science Foundation grant EIA–9876619 to the National Institute of Statistical Sciences, and by the National Center for Education Statistics.

way that intruders cannot compromise the privacy of data subjects. As data disseminators, government agencies wish to provide as much information as possible to the data user while satisfying the mandate of confidentiality.

Various methods have been proposed for protecting confidentiality (of data) and privacy (of data subjects). These methods can be broadly classified as methods that are geared to the release of microdata records—for example, aggregation, data swapping, jittering, variable (attribute) and cell suppression, and those that deal with release of data summaries—for example, intelligent table servers [8, 11] and regression servers (that selectively release information requested by user queries).

Data swapping achieves confidentiality protection by selectively modifying a fraction of records, thus making it impossible for an intruder to be certain that any record in the microdata corresponds to an actual data subject. Dalenius and Reiss [7] appear to have been the first to use the term "data swapping." They present theoretical results indicating that, under various assumptions, it should be possible to preserve statistics defined on any specified dimension of the microdata. Reiss [16] presents a modified technique called "approximate data swapping" that simulates data from the joint distribution that one is interested in maintaining. However, this technique appears closer to synthetic data generation than to what is generally understood as data swapping. Zayatz et al. [20] discuss data swapping in the context of the 2000 Census. Boyd and Vickers [4] refer to "record swapping" when swapping under constraints (on unswapped attributes). Moore[13, 14] discusses "rank swapping" and the "confidentiality edit," which are also versions of constrained data swapping (see §2 below for details).

Clearly data swapping alters the data. While this change reduces the risk of violating confidentiality, it may also diminish the utility of the data to users. In this paper we study effects of data swapping through a number of distortion measures of (dis-)utility, comparing their behavior as the swapping rate and swapped attributes are varied. Winkler [18] and Yancey et al. [19] have studied risk measures for data swapping, primarily from the point of re-identification of records. Domingo-Ferrer et al. [9] have applied various measures of risk and utility to U.S. Census data.

We lay out basic notation and terminology in §2. §3 discusses measures of distortion that can be used to characterize the effects of swapping. Data from the Current Population Survey are used for illustrative purposes throughout this section. §4 contains a discussion.

Table 1: *Example of microdata with information on six records for average weekly work hours, employer type, sex, and marital status.*

Rec. No.	AvgHrs	EmpTyp	Sex	MarStat
1	< 40	Gov	M	M
2	40	SelfEmp	F	UM
3	< 40	Priv	F	M
4	> 40	Priv	M	M
5	> 40	SelfEmp	F	UM
6	40	Oth	F	M

2 Basic Terminology

Conceptualize the microdata or data file as a matrix, with the rows (also called records) representing individuals or observations, and columns representing variables (attributes) for which information on the observations is collected. Let N be the number of records, and let v be the total number of variables in the microdata. Denote the j^{th} variable in the microdata by X_j, and the i^{th} record by $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iv})$.

A formal definition of data swapping is given in Willenborg and de Waal [17], who define a data swap of $2k$ elements in terms of k elementary swaps. An *elementary swap* is a random selection of two records i and j from the microdata and an interchange of the values of variables being swapped for these two records. An informal definition of data swapping, that helps visualize the process, is given by Duncan and Keller-McNulty [10] who have referred to data swapping as "switching column values for pairs of rows." Thus, if we are swapping the values of X_1 for records i and j, the post-swap values of the i^{th} and j^{th} records will be $(x_{j1}, x_{i2}, \ldots, x_{iv})$ and $(x_{i1}, x_{j2}, \ldots, x_{jv})$, respectively. As Willenborg and de Waal [17] point out, one way of carrying out k elementary swaps is to pick $2k$ random numbers from 1 to N without replacement, and interchange values of the variable(s) to be swapped for the records corresponding to the $(2i-1)^{th}$ and the $2i^{th}$ draws, $i = 1, \ldots, k$. When the candidates for a swap pair are picked at random we will refer to the resulting swaps as *random swaps*. By default we assume that elements of a swap pair are picked without replacement—thus no record appears in more than one swap pair.

Clearly not all variables in a record are swapped. We call the subset of variables that will be swapped the *swap variables* or *swapped variables*. The fraction of the total N records in the microdata (i.e., $2k/N$, if all k elementary swaps were implemented) that are swapped is called the *swap proportion* or *rate*. Henceforth, when we swap appropriate fields between records i and j, we use (i, j) to denote the *swap pair*.

In some situations there may be conditions on pairs of records, defined by variables other than swap variables, in order for the two records to be feasible swap candidates. For instance, for the data given in Table 1 we may only allow swaps between records with the same value for *Sex*. In this case the swap pair (2,3) is feasible, whereas (1,3) is not. If we prohibit swaps for records with the same value of *MarStat*, (2,3) is a feasible swap, whereas (3,4) is not. Variables whose values define the feasibility of a swap pair will be called *constraining variables*. We emphasize that constraining variables *are not swap variables*. The simplest constraints, as exemplified above, are those of exact equality or inequality of the constraining variable(s) associated with the pair of records to be swapped.

Let s be the number of swap variables, and c be the number of constraining variables. We can partition the collection of v variables as $\mathbf{X} = (\mathbf{X}^S, \mathbf{X}^C, \mathbf{X}^U)$ where $\mathbf{X}^S$ denotes the swap variables, $\mathbf{X}^C$ denotes the constraining variables, and $\mathbf{X}^U$ denotes the variables that are neither swap nor constraining variables. We sometimes refer to the post-swap values as $\mathbf{Y} = (\mathbf{Y}^S, \mathbf{Y}^C, \mathbf{Y}^U)$. (Note: $(\mathbf{X}^C, \mathbf{X}^U) \equiv (\mathbf{Y}^C, \mathbf{Y}^U)$.)

When more than one variable in a data set is to be swapped, there are different ways to effect the swap. We assume that swapped variables are swapped *simultaneously*: if we are swapping $2k$ records on $s = 2$ variables X_1^S and X_2^S, then we pick k pairs of records and swap the values of both X_1^S and X_2^S for the two records in a pair. This method preserves relationships among the swapped variables, but modifies the relationship of swapped variables to other variables in the data. (It is also possible to swap *sequentially* by carrying out s single-variable swaps in sequence. In this case, for $s = 2$, we first pick k pairs of records and do k elementary swaps for the values of X_1^S, and then independently pick another k pairs and swap the values of X_2^S.)

Random swaps may not always result in different values for the swapped variables. If we swap *AvgHrs* for the pair (1,2) the records for these two observations are different pre- and post-swap, whereas if we swap *AvgHrs* for the pair (1,3) both records remain unchanged. We call the first case a *true swap*, i.e., a true swap results in different pre- and post-swap values for the records being swapped. Conversely, a swap that results in no change in record values is called a *false swap*.

When a subset of the collection of swaps result in the same pre- and post-swap frequency distribution for the records, then the swaps are said to be *compensating swaps*. For example, for the data given in Table 1, if $k = 1$ and the pair swapped is $(2, 5)$, then the frequency of the combinations (F, UM, 40, SelfEmp) and that of (F, UM, $>$ 40, SelfEmp) in the microdata does not change. More complicated examples, where compensation occurs even when the values of the non-swap variables are not equal for every swap pair, can be constructed.

3 Measures of Distortion

Inevitably, data confidentiality measures distort joint distributions in the data. One advantage of data swapping is that, as only switches of values between records are involved, the univariate marginal distributions of all variables in the microdata are preserved. In addition, because (in our setting) when multiple variables are swapped they are swapped simultaneously, joint distributions not involving swap variables or only involving swap variables are also preserved. However, joint distributions that involve both swap and non-swap variables can be distorted. Dalenius and Reiss [7] have theoretical results that imply the possibility of preserving p-variate statistics in the swapped data ($p \leq v$), but it is not clear that the assumptions under which their results are derived hold in practical situations.

To measure changes in joint distributions, we consider five different measures of distortion between pre- and post-swap data. Hellinger distance and total variation distance are standard measures of distance between distributions, Cramer's V and the contingency coefficient C are measures of association for bivariate distributions (more specifically, $m \times n$ contingency tables), and entropy is an information-theoretic measure of uncertainty. Total variation distance for univariate distributions, entropy-based measures, Cramer's V, and the contingency coefficient have been considered by other authors (see subsections that follow) in the context of data swapping. However, Hellinger distance does not appear to have been used previously, nor has the general form of total variation distance.

We study the effect of swap variables and swap rates on data from the Current Population Survey (CPS), using a modified version of the CPS for 1993 obtained from [1]. This version has 48,842 observations and retains only 8 variables, some of whose values have been aggregated. We will refer to this data set as CPS-8d data. The variables present in the data and definitions of their categories are given in Table 3.

We studied the effect of three swap proportions (0.01, 0.05, and 0.1) on unconstrained swaps of a single variable at a time, i.e., $s = 1$, $c = 0$, and $u = 7$. However, as we swap every one of the eight variables, we will be considering multiple permutations of $\mathbf{X}$, i.e., X_1 will represent each of the variables in the data set in turn. When $\mathbf{X}$ (as well as its observed value $\mathbf{x}$) represents the entire data vector, we are measuring the distance between the v-variate joint distributions in the data. However, in some cases we may be interested in either preserving (or destroying) some lower-dimensional joint distribution. Replacing $\mathbf{X}$ by the appropriate

Table 2: *Variables and Category Values for CPS-8d data.*

Variable Name	Categories
Age (in years)	<25, 25–55, >55
Employer Type	Govt., Priv., Self-Emp., Other
Education	<HS, HS, Bach, Bach+, Coll
Marital Status	Married, Other
Race	White, Non-White
Sex	Male, Female
Average Weekly Hours Worked	< 40, 40, > 40
Annual Salary	<\$50K, \$50K+

v_1 dimensional vector ($v_1 \leq v$) representing the variables in the distribution of interest allows us to compute the distortion for the corresponding v_1 dimensional joint distribution. In the following subsections we will denote the empirical density present in the pre-swap data as f, and that in the post-swap data as g.

We computed Hellinger distance, total variation distance, and change in entropy for the complete 8-variate distribution. As measures derived from Cramer's V and the contingency coefficient C were computable only for all bivariate distributions in the data, we computed all bivariate Hellinger distances, total variation distances, and entropy changes for comparison.

3.1 Hellinger Distance

Hellinger distance ([12], for example) between distributions f and g on a countable state space is defined as

$$H(f,g) = \frac{1}{\sqrt{2}} \sqrt{\sum_{\mathbf{x}} \left(\sqrt{f(\mathbf{x})} - \sqrt{g(\mathbf{x})} \right)^2}. \tag{1}$$

We note that the same absolute difference in $f(\mathbf{x})$ and $g(\mathbf{x})$ affects the Hellinger distance to a greater extent when the value of $f(\mathbf{x})$ is small. The interpretation of $H(f,g)$ is as the sine of the angle between the Hilbert vectors representing $\sqrt{f}$ and $\sqrt{g}$. Each square-root density can itself be interpreted as a point on the unit sphere in a real Hilbert space. Hellinger distance also corresponds to Cressie-Read divergence (see [5, 6]) with $\lambda = -0.5$.

Figure 3.1 plots 8-way Hellinger distance for all three rates, and Table 3 gives 2-way Hellinger distances for the 5% swapping rate. From Figure 3.1 we see that distortion increases as swap proportion increases. The minimum distortion across all three rates for

Table 3: *Values of 2-way Hellinger distance and averages for fixed swap variable (in column margin) for 5% swapping proportion. Column variable is swapped, and row variable is the other variable in pair for adV computation. Row maxima are indicated by* **bold face** *and minima by italics, as are maxima and minima for the mean.*

	Age	EmpTyp	Edu	MS	Race	Sex	AvgHrs	AnnSal
Age	–	0.0763	0.0730	0.0900	*0.0440*	0.0511	0.0649	**0.1193**
EmpTyp	**0.0740**	–	*0.0476*	0.0583	0.0652	0.0595	0.0592	0.0548
Edu	0.0882	0.0613	–	0.0450	0.0468	*0.0386*	0.0518	**0.0934**
MS	0.0912	0.0585	*0.0405*	–	0.0566	0.0990	0.0550	**0.1167**
Race	0.0357	0.0468	*0.0173*	0.0440	–	0.0471	0.0266	**0.0472**
Sex	0.0435	0.0532	*0.0364*	**0.0939**	0.0557	–	0.0631	0.0800
AvgHrs	0.0810	0.0649	*0.0508*	0.0635	0.0592	0.0787	–	**0.0819**
AnnSal	**0.1061**	*0.0524*	0.0638	0.0982	0.0547	0.0711	0.0591	–
Mean	0.0742	0.0591	*0.0471*	0.0704	0.0546	0.0636	0.0542	**0.0848**

8-way Hellinger distance is obtained when *Edu* or *AvgHrs* is swapped, and the maximum when *AnnSal* is swapped. From Table 3 we see that for the 5% proportion a swap of *AnnSal* is most likely to maximize distortion, whereas a swap of *Edu* often results in minimum distortion. This holds also for the mean distortion, where the average is taken over non-swap variables in the bivariate distribution.

3.2 Total Variation Distance

For a countable state space the definition of total variation distance is

$$TV(f,g) = \frac{1}{2}\sum_{\mathbf{x}} |f(\mathbf{x}) - g(\mathbf{x})|\,.$$

The "index of dissimilarity" considered by Moore [14] is a special case of this distance defined for a single variable.

As for Hellinger distance, from Figure 3.1 we see that total variation distance in the 8-way distribution increases as the swapping rate increases. *Race* or *Edu* are the minimizers of 8-way total variation distance, *AnnSal* or *MS* are its maximizers.

From Table 4, which contains total variation distances for bivariate distributions for the 5% swap rate, we see that *Edu* and *Race* are the most frequent minimizers. *MS* dominates

Figure 1: *Graph of 8-way Hellinger and total variation distances, and 8-way entropy change for 1% swap proportion (circles), 5% swap proportion (triangles) and 10% swap proportion (pluses).*

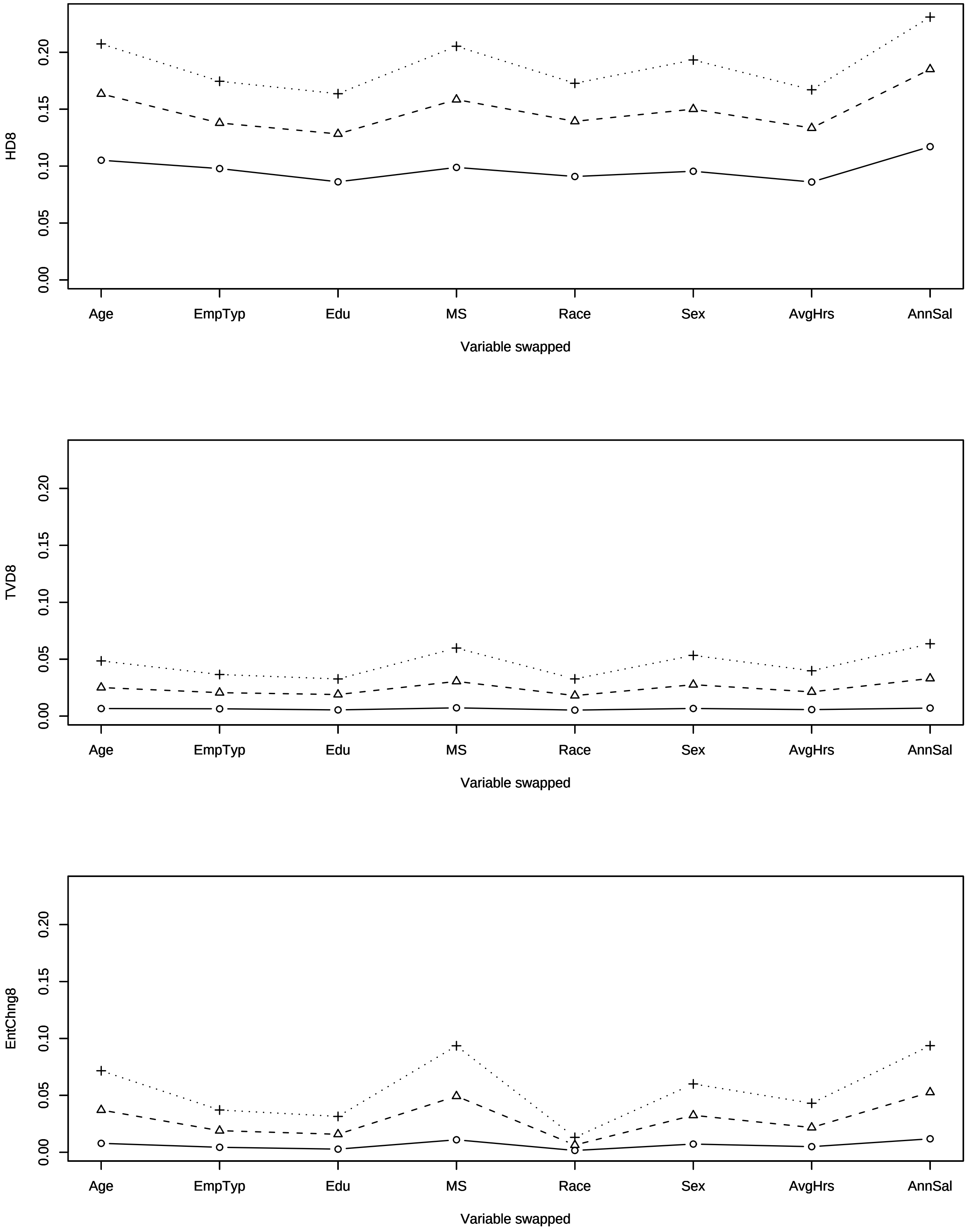

Table 4: *Values of 2-way total variation distance and averages for fixed swap variable (in column margin) for 5% swapping proportion. Column variable is swapped, and row variable is the other variable in pair for adV computation. Row maxima are indicated by* **bold face** *and minima by italics, as are maxima and minima for the mean.*

	Age	EmpTyp	Edu	MS	Race	Sex	AvgHrs	AnnSal
Age	–	0.0070	0.0067	**0.0112**	*0.0028*	0.0054	0.0086	0.0100
EmpTyp	**0.0071**	–	*0.0036*	0.0062	0.0041	0.0045	0.0060	0.0057
Edu	0.0103	0.0066	–	0.0049	0.0038	*0.0030*	0.0060	**0.0160**
MS	0.0116	0.0067	*0.0037*	–	0.0062	0.0218	0.0060	**0.0257**
Race	0.0023	0.0021	*0.0003*	0.0037	–	**0.0043**	0.0011	0.0033
Sex	0.0040	0.0037	*0.0027*	**0.0195**	0.0061	–	0.0074	0.0120
AvgHrs	**0.0132**	0.0073	0.0060	0.0078	*0.0055*	0.0115	–	0.0117
AnnSal	0.0074	0.0046	0.0075	**0.0179**	*0.0045*	0.0095	0.0061	–
Mean	0.0080	0.0054	*0.0043*	0.0102	0.0047	0.0086	0.0059	**0.0121**

as a maximizer, with *AnnSal* and *Age* as close seconds. However, *AnnSal* as a swap variable maximizes the average total variation distance, and *Edu* minimizes it.

3.3 Change in Entropy

The usual formula for Shannon entropy is given by

$$\sum_{\mathbf{x}} f(\mathbf{x}) \log\left(f(\mathbf{x})\right)$$

in our notation. The formula leads to the natural interpretation of Shannon entropy as the expectation of a random variable that takes values $\log(f(\mathbf{X}))$ with probability $f(\mathbf{X})$. An alternative interpretation is that it is "the minimum average number of "yes or no" questions required to determine the result of one observation of $\mathbf{X}$" [3].

This entropy function takes its largest value when all possible values of $\mathbf{X}$ have the same probability of being observed, and the smallest when all of the probability mass is concentrated on a single value.

Shannon entropy has been considered as a distortion measure by other authors (see [17]). However, they consider conditional entropy, whereas we use use post-swap entropy minus

Table 5: *Values of 2-way entropy change and averages for fixed swap variable (in column margin) for 5% swapping proportion. Column variable is swapped, and row variable is the other variable in pair for adV computation. Row maxima are indicated by* **bold face** *and minima by italics, as are maxima and minima for the mean.*

	Age	EmpTyp	Edu	MS	Race	Sex	AvgHrs	AnnSal
Age	–	0.0074	0.0059	0.0133	*0.0003*	0.0016	0.0073	**0.0138**
EmpTyp	**0.0072**	–	0.0024	0.0030	*0.0018*	0.0027	0.0044	0.0023
Edu	0.0086	0.0042	–	0.0013	*0.0005*	*0.0005*	0.0030	**0.0145**
MS	0.0136	0.0030	*0.0011*	–	0.0016	0.0208	0.0039	**0.0282**
Race	0.0002	0.0010	*0.0000*	0.0010	–	**0.0012**	0.0004	0.0010
Sex	0.0012	0.0022	*0.0005*	**0.0188**	0.0016	–	0.0056	0.0070
AvgHrs	**0.0110**	0.0052	0.0029	0.0051	*0.0019*	0.0084	–	0.0089
AnnSal	**0.0108**	0.0021	0.0068	0.0203	*0.0013*	0.0056	0.0048	–
Mean	0.0075	0.0036	0.0028	0.0090	*0.0013*	0.0058	0.0042	**0.0108**

pre-swap entropy to quantify entropy change (EC). That is,

$$EC = \sum_{\mathbf{x}} g(\mathbf{x}) \log\left(g(\mathbf{x})\right) - \sum_{\mathbf{x}} f(\mathbf{x}) \log\left(f(\mathbf{x})\right).$$

Positive values of EC indicate that swapping has increased the uncertainty in the data.

We note from Figure 3.1 that entropy change in the 8-way distribution increases as the swap proportion increases. The ordering of swap variables in terms of distortion produced in the 8-way distribution is exactly the same over all three swap rates. *AnnSal* and *MS* are maximizers of distortion, whereas *Race* and *Edu* are its minimizers.

Table 5 gives values of the post-swap bivariate entropy minus the pre-swap bivariate entropy. *AnnSal* and *Age* appear to be the dominant maximizers of the entropy change here, and *Race* and *Edu* are the predominant minimizers. Average entropy change is minimized by *Race*, with *Edu* in second place, and maximized by *AnnSal*.

3.4 Measure Based on Cramer's V

Cramer's V is a measure of association based on the χ^2 statistic for a $m \times n$ contingency table. It is defined as

$$V = \sqrt{\frac{\chi^2}{N \min(m-1, n-1)}}, \tag{2}$$

where χ^2 is the usual χ^2 defined for the test of independence. For 2×2 tables the square of Cramer's V simplifies to a measure called *phi* and equals Goodman and Kruskal's tau [2, 15]. Cramer's V lies between 0 and 1—a value of 0 indicates no association, whereas a value of 1 indicates perfect association. It is more difficult to interpret values between the extremes.

Cramer's V has been used by Boyd and Vickers [4] in the context of data swapping. However, they used Cramer's V on pre and post values of swap variables (within geographical subsets of the swapped population) to assess the effect of the swap. Use of the measure in this fashion for the entire data set would amount to quantifying some equivalent of the swap rate.

In order to measure distortion due to swapping for any bivariate distribution we define

$$adV_{ij} = V_{ij}^{pre} - V_{ij}^{post},$$

where V_{ij}^{pre} is Cramer's V defined for the cross table obtained from X_i and X_j, where $i = 2, \ldots, v$, $j = 1$ (Recall that the variables are permuted so that j represents each of the variables in the data in turn). V_{ij}^{post} is defined on the post-swap cross table $Y_i \times Y_j$. Like Cramer's V, adV_{ij} ranges from 0 to 1. Positive values of adV_{ij} indicate that swapping has weakened the association between X_i and X_j.

The behavior of adV was similar for all three rates studied. The range of adV values for the 1%, 5%, and 10% rates were approximately 0.001–0.012, 0.005–0.046, and 0.008–0.120, indicating an approximately linear scaling of the distortion with respect to swap proportions in the range 0.01–0.10. In Table 6 we present the values of adV for the 5% swap. *AnnSal* is a dominant maximizer of adV, and *Edu* is a dominant minimizer. The highest average distortion is due to *AnnSal*, whereas the lowest is due to *Edu*.

3.5 Measure Based on Contingency Coefficient C

Pearson's contingency coefficient C also measures association, and is based on the χ^2. For an $m \times n$ contingency table it is defined as

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}, \tag{3}$$

where χ^2 is the usual χ^2 defined for the test of independence. Like Cramer's V, C lies between 0 and 1. However, its upper limit depends on m and n, and it is difficult to compare tables of different sizes with this measure. Like Cramer's V, it also suffers from the difficulty of interpretation for intermediate values.

As for Cramer's V we define

$$adC_{ij} = C_{ij}^{pre} - C_{ij}^{post},$$

Table 6: *Values of 2-way adV and averages for fixed swap variable (in column margin) for 5% swapping proportion. Column variable is swapped, and row variable is the other variable in pair for adV computation. Row maxima are indicated by* **bold face** *and minima by italics, as are maxima and minima for the mean.*

	Age	EmpTyp	Edu	MS	Race	Sex	AvgHrs	AnnSal
Age	–	0.0190	0.0129	0.0305	*0.0103*	0.0150	0.0166	**0.0308**
EmpTyp	**0.0190**	–	*0.0067*	0.0180	0.0183	0.0167	0.0141	0.0164
Edu	0.0196	0.0121	–	0.0106	0.0111	*0.0077*	0.0108	**0.0445**
MS	0.0313	0.0180	*0.0086*	–	0.0176	0.0463	0.0160	**0.0603**
Race	0.0053	0.0095	*0.0001*	0.0106	–	**0.0131**	0.0034	0.0110
Sex	0.0112	0.0130	*0.0074*	**0.0416**	0.0183	–	0.0212	0.0300
AvgHrs	0.0250	0.0174	*0.0101*	0.0214	0.0172	0.0325	–	**0.0336**
AnnSal	0.0228	*0.0145*	0.0197	**0.0420**	0.0149	0.0235	0.0174	–
Mean	0.0192	0.0148	*0.0094*	0.0250	0.0154	0.0221	0.0142	**0.0324**

where C_{ij}^{pre} is the contingency coefficient defined for the cross table obtained from X_i and X_j, where $i = 2, \ldots, v$, $j = 1$. The cross table obtained from Y_i and Y_j is used to define C_{ij}^{post}. Observed values of adC_{ij} for the CPS-8d data range from 0 to 1. Positive values of adC_{ij} indicate that swapping has weakened the association between X_i and X_j. Comparison of adC_{ij} values across cross tables that have different sizes is, as noted above, problematic.

Table 7 shows how this measure performs for the 5% swap. We see that the performance of *adC* is very similar to that of *adV* with almost the same minimizers and maximizers. We caution against comparisons within this table as the bivariate distributions have different dimensions.

3.6 Comparison of Performance

Overall there is significant consistency in the conclusions drawn from the different measures considered. In all cases, distortion increases as the swap rate increases, with 8-way Hellinger distance showing increase approximately proportionate to the swap rate. For the complete 8-way distribution total variation distance and entropy change result in almost the same ordering of swap variables in terms of resulting distortion. Hellinger distance shows somewhat different ordering (*Age* and *AvgHrs* appear to contribute the most to this change in order).

For bivariate distributions *Age* and *AnnSal* are the preponderant maximizers for both

Table 7: *Values of 2-way adC and averages for fixed swap variable (in column margin) for 5% swapping proportion. Column variable is swapped, and row variable is the other variable in pair for adV computation. Row maxima are indicated by* **bold face** *and minima by italics, as are maxima and minima for the mean.*

	Age	EmpTyp	Edu	MS	Race	Sex	AvgHrs	AnnSal
Age	–	0.0245	0.0166	0.0259	*0.0103*	0.0148	0.0202	**0.0286**
EmpTyp	**0.0244**	–	*0.0109*	0.0173	0.0181	0.0163	0.0184	0.0159
Edu	0.0253	0.0196	–	0.0104	0.0110	*0.0076*	0.0143	**0.0382**
MS	0.0266	0.0173	*0.0085*	–	0.0174	0.0372	0.0148	**0.0479**
Race	0.0053	0.0094	*0.0001*	0.0105	–	**0.0129**	0.0034	0.0109
Sex	0.0110	0.0126	*0.0074*	**0.0333**	0.0181	–	0.0192	0.0283
AvgHrs	0.0306	0.0227	*0.0134*	0.0199	0.0169	0.0297	–	**0.0307**
AnnSal	0.0211	*0.0141*	0.0167	**0.0331**	0.0147	0.0222	0.0158	–
Mean	0.0206	0.0172	*0.0105*	0.0215	0.0152	0.0201	0.0152	**0.0286**

Hellinger distance and entropy change, whereas *MS* also plays a significant role for total variation distance. *Race* and *Edu* are the primary minimizers for all three of these measures. The behavior of *adV* and *adC* is primarily like that of 2-way Hellinger distance in that *AnnSal* is most likely to be a maximizer and *Edu* is most likely to be a minimizer. Interestingly, however, *MS* plays a stronger role than *Age* in maximizing *adC* and *adV*. For all 2-way measures, the average distortion is maximized by *AnnSal.* For most 2-way measures *Edu* is a minimizer of average distortion, the exception being 2-way entropy change where *Race* minimizes distortion over *Edu.*

Thus, it is clear that there are variables whose swapping leads consistently to higher distortions, while others show consistently low distortions. Furthermore, when considering extremes of average distortion, all of the other measures behave like Hellinger distance.

4 Discussion

Various measures of distortion that quantify changes in joint distributions have been studied. While we have studied the performance of these measures on the joint distributions present in the entire microdata, they can be used to look at subsets of the data.

Some of the measures that we have looked at have been considered in the context of data swapping but not in the same fashion. For instance, Boyd and Vickers [4] have used Cramer's V and the contingency coefficient C. Moore [14] has considered the index of dissimilarity, which is a special case of total variation distance. Of the measures considered, the ones based on Cramer's V and the contingency coefficient C are defined only for bivariate distributions. Hellinger distance, total variation distance, and change in entropy have the advantage that they are defined for higher dimensional distributions too.

The measures presented here are particularly suited to categorical data. The primary characterization that unifies them is that they utilize only cell counts or frequencies. They are not suited for continuous-valued variables, where all observed frequencies are close to 0. Alternative measures or strategies must be devised for such variables. For ordinal or interval scale data one might want to use a measure that also incorporates the value of the swapped entity. For example, one might penalize data shifts to values further away from the initial value more severely.

References

[1] Current population survey. *http://www.bls.census.gov/cps/cpsmain.htm.*

[2] A. Agresti. *Categorical Data Analysis.* Wiley, New York, 1990.

[3] R. B. Ash. *Information Theory.* Dover Publication, Inc., New York, 1965.

[4] M. Boyd and P. Vickers. Record swapping - a possible disclosure control approach for the 2001 UK Census. *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, 1999.

[5] N. A. C. Cressie and T. R. C. Read. Cressie-read statistic. In S. Kotz and N. L. Johnson, editors, *Encyclopedia of Statistical Sciences, Supplementary Volume.* Wiley, New York, 1988.

[6] N. A. C. Cressie and T. R. C. Read. *Goodness-of-Fit Statistics for Discrete Multivariate Data.* Springer–Verlag, New York, 1988.

[7] T. Dalenius and S. P. Reiss. Data-swapping: A technique for disclosure limitation. *J. Statist. Planning Inf.*, 6:73–85, 1982.

[8] A. Dobra, A. F. Karr, S. E. Fienberg, and A. P. Sanil. Software systems for tabular data releases. *Int. J. Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):529–544, 2002.

[9] J. Domingo-Ferrer, J. M. Mateo-Sanz, and V. Torra. Comparing SDC methods for microdata on the basis of information loss and disclosure risk. *presented at UNECE Workshop on Statistical Data Editing*, May, 2001.

[10] G. T. Duncan and S. A. Keller-McNulty. The impact of data swapping on confidentiality and data utility. *Talk presented to Institute for Social Research, University of Michigan*, 2000.

[11] A. F. Karr, J. Lee, A. P. Sanil, J. Hernandez, S. Karimi, and K. Litwin. Disseminating information but protecting confidentiality. *IEEE Computer*, 34(2):36–37, 2001.

[12] L. Le Cam and G. L. Yang. *Asymptotics in Statistics.* Springer–Verlag, New York, 1990.

[13] R. A. Moore. Controlled data-swapping techniques for masked public use microdata sets. *U. S. Bureau of the Census, Statistical Research Division, Washington D.C.*, 1996.

[14] R. A. Moore. Preliminary recommendations for disclosure limitation for the 2000 census: Improving the 1990 confidentiality edit procedure. *U. S. Bureau of the Census, Statistical Research Division, Washington D.C.*, 1996.

[15] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C.* Cambridge University Press, New York, 1992.

[16] S. P. Reiss. Practical data-swapping: The first steps. *ACM Trans. Database Systems*, 9(1):20–37, 1984.

[17] L. C. R. J. Willenborg and T. de Waal. *Elements of Statistical Disclosure Limitation.* Springer–Verlag, New York, 2000.

[18] W. E. Winkler. Re-identification methods for evaluating the confidentiality of analytically valid microdata. *Res. Official Statist.*, 1:87–104, 1998.

[19] W. E. Yancey, W. E. Winkler, and R. H. Creecy. Disclosure risk assessment in perturbative microdata. *Inf. Control in Statist. Databases*, 2002.

[20] L. Zayatz, P. Steel, and S. Rowland. Disclosure limitation for census 2000. *U. S. Bureau of the Census, Statistical Research Division, Washington D.C.*, 2000.