RESEARCH REPORT SERIES
(Statistics #2007-19)

Analytically Valid Discrete Microdata Files and Re-identification

William E. Winkler

Statistical Research Division
U.S. Census Bureau
Washington, DC 20233

Report Issued: December 10, 2007

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

Analytically Valid Discrete Microdata Files and Re-identification

William E. Winkler[1], william.e.winkler@census.gov
U.S. Census Bureau, Washington, DC 20233-9100

Abstract

Loglinear modeling methods have become quite straightforward to apply to discrete data **X**. A good-fitting loglinear model can be used to generate synthetic copies of **X1**, …, **Xn** of **X** that preserve analytic properties but may allow re-identification of small cells. With fitting algorithms that use more general convex constraints and are designed to deal with missing data, we are able to disperse the counts associated with small cells over other cells in a manner that reduces re-identification risk while still maintaining most analytic properties.

Keywords: Data Quality, Loglinear Model Fit, Missing Data, Convex Constraints

1. Introduction

A primary purpose of collecting data is to produce a data file that can be used for two or more analytic purposes. The methods for 'cleaning up' the data include edit/imputation for removing 'implausible combinations of values of fields' and for imputing values that are consistent with underlying uses of the data. The intention of the 'clean-up' procedures is to produce data that are 'fit for use.' In an ideal world, a final data file that has gone through extensive 'clean-up' could be used for several (or all) sets of analyses.

Original, 'cleaned-up' microdata are considered much more useful for analytic purposes, particularly ad hoc ones, than tabulations from microdata that are in publications. There is an extensive literature on methods for producing public-use microdata. The methods are subdivided into two subsets: (1) those that are easy to implement (typically performed in statistical agencies) and (2) those that are more difficult to implement due to the need for greater modeling skills and sophisticated programming algorithms. Few, if any, of the easy-to-implement methods have been justified in terms of producing analytically valid data (Winkler 2004, 2007b). By an *analytically valid* file, we mean a file that will allow the approximate reproduction of two or more sets of analyses that could be produced from original data.

If the data are discrete and there are a moderate number of variables (10-20), then it is straightforward to apply loglinear modeling methods to obtain a good fitting model. If synthetic data are drawn from the model, then most analytic properties will be maintained in the synthetic data. For the analytic properties, we are most concerned with the larger cells and largest margins. If small cells are varied slightly in a manner that does not substantially distort most margins, then the analytic properties may not change much. We make the assumption that small cells (say those having counts below three) are where re-identification is easiest.

We need an overall modeling/edit/imputation framework that includes that standard loglinear methods, standard methods for dealing with missing data (Little and Rubin 2002), methods of dealing with structural zeros, and methods of general convex constraints. *Structural zeros* are cells that edit restraints force to have counts of zero. As an example, we do not want to have cells for which a child of less than 15 is married. In creating the framework, Winkler (2007a) used general discrete editing methods (Winkler 1997) to create a structure into which the missing-data-modeling methods of Little and Rubin (2002) could be imbedded. Winkler (1993) showed that similar methods and algorithms for unsupervised learning of mixture models under a variety of constraints could be used for record linkage. D'Orazio, Di Zio, and Scanu (2006) demonstrated that the basic algorithms of Winkler (1993) could also be used for statistical matching under convex constraints. The framework can be used for creating better quality data that meets a variety of analytic constraints (Winkler 2003, 2007a; D'Orazio, Di Zio, and Scanu 2006). In this paper we use the framework for the special case of creating the desired models from which synthetic data can be created. If we use these generalized methods and software for the creation, clean-up, and understanding of the original non-public data **X**, then it is quite straightforward to use the methods and software for the creation of the models for the synthetic data.

We create a model **M2** from which we can draw synthetic data **X2** in several steps. In the first step, we merely create a model **M** with the set of key interactions that fits the original data **X** well. Synthetic data **X0** drawn from **M** will likely approximately reproduce many small cells. The next model **M1** uses the same interactions as model **M** but 'disperses' the observed counts from the small cells across both the small cells and the sampling zero cells. A cell that is zero in an observed file (sample) but may be nonzero in a large corresponding population file is referred to as a *sampling zero*. In creating **M1**, we use standard missing data methods (Little and Rubin 2002) that keep the total count for the set of small cells and sampling zero cells approximately fixed. This is intended to better preserve overall analytic properties. Because resultant model **M1** (from the final fitted solution) may still yield synthetic data **X1** that allows some re-identification, we fit a second model that has convex constraints that control the size of the fitted counts in the final model **M2**. The convex constraints can allow us to specify upper and lower bounds on the originally observed small cells and the original sampling zeros. The model **M2** still preserves overall analytic properties but significantly reduces the risk of re-identification with any synthetic data **X2** that is drawn from **M2**.

The outline of this paper is as follows. Following this introduction, we provide background on basic loglinear modeling and missing data methods and their extensions for general edit/imputation. In the third (data and methods) section, we describe the extended loglinear methods for modeling with discrete data and how we can use convex constraints to restrict the upper and lower bounds in certain cells. In the fourth section, we provide results demonstrating that the new methods yield models and synthetic data that quite accurately reproduce properties of the original, confidential data while reducing re-identification risk. The fifth section consists of discussion and the sixth section is concluding remarks.

2. Background

In this section we provide background on methods of creating loglinear model **M** (Bishop, Fienberg and Holland 1975) that are straightforward to apply to general discrete data. We also provide background on general methods of imputation and editing for missing data under convex constraints that extend the basic methods and can also be straightforward to apply. By *straightforward to apply* we mean that the general methods and software can be applied without any modifications that are specific to a particular data file or analytic use.

3.1 Loglinear Modeling and Creating Synthetic Data

Standard references for loglinear modeling and categorical data analysis are Agresti (2007) and Bishop, Fienberg, and Holland (1975). If $\mathbf{X_i}$ is a table of cell counts from a discrete population, then we are interested in finding a more parsimonious representation $\hat{X}_i$ of the table. If table $\mathbf{X_i}$ has n = I × J × K cells, then we fit models via an iterative proportional fitting procedure (IPF) in which we fit to specific observed margins in cyclic order. In this case, the general index **i** = (i, j, k). The procedure is assured to increase likelihood and typically increases to a maximum likelihood estimate under a multinomial or Poisson model. We will assume multinomial.

The original table has I × J × K-1 degrees of freedom. The different fitted models will have fewer degrees of freedom. If we fit an independence model, then we successively fit to the observed margins for each single variable and repeat until convergence. If we fit using an all 2-way interaction model, we fit successively to the set of margins determined by I × J, then I × K, then J × K and repeat in a cyclic manner. In the original table $\mathbf{X_i}$ we can have sampling or structural zeros. A structural zero must stay at zero in any fitted model (Winkler 1990) whereas a sampling zero can become non-zero in the fitted model. Instead of using counts for $\mathbf{X_i}$, we might divide every cell count by the total to obtain a probability.

If the original table $\mathbf{X_i}$ has many cells with small counts, then we would expect any good-fitting model $\hat{X}_i$ to also have many cells with small counts. If we draw a set of 'synthetic' records from the model $\hat{X}_i$, then we expect that the synthetic data would allow reproduction of many analyses. Because small counts may yield a few re-identifications, the statistical agency may have a policy that every nonzero cell must have a count above a lower bound (say 3 or 5). To a fitted model that that satisfies the basic interaction model and also lower bounds of cells, we might use convex constraints on complementary cells. If the set of convex constraints is consistent with the interaction model and the set of structural zeros, then we would necessarily have a solution (Winkler 1990).

As an example, assume that I, J, K each take values 0 and 1 and that we have fit a 2-way interaction model. In obtaining $\hat{X}_i$, successively fit to the I × J, I × K, and J × K margins. If $\hat{P}$ (1,1,1), the probability associated with fitted cell (1,1,1), must be above a lower bound, then we can place restrictions (convex upper bound constraints) on $\hat{P}$ (1,1,0), $\hat{P}$ (1,0,1), and $\hat{P}$ (0,1,1) because $\hat{P}$ (1,1,1)+ $\hat{P}$ (1,1,0) is fitted to the observed margin P(1,1,1)+P(1,1,0), and so on. If certain fitted cells cannot be forced above an upper bound, then we might use a slightly ad hoc procedure of setting the fitted cell to zero and forcing certain complementary cells upward so that we still have a probability measure. The synthetic data can be drawn from the resultant model. We observe that the procedures allow us to place upper and lower bounds on the fitted values associated with certain cells or combinations of cells.

2.2 Missing Data Imputation, Editing, and Convex Constraints

In this section we describe standard methods of creating a loglinear model **Y** (Bishop, Fienberg and Holland 1975) that is straightforward to extend to models that take account of *edit restraints* (Fellegi and Holt 1976). If **X** is original data and $\mathbf{X_i}$ is a specific cell, then $\mathbf{X_i}$ is a structural zero or cell forbidden by an edit if its count must be zero. With discrete data, $\mathbf{X_i}$ might be associated with a child of less than 16 years of age in a household who is married or a child of less than 22 who has a college postgraduate degree. The theoretical justification connecting editing (structural zeros) with imputation as in Little and Rubin 2002 or general loglinear modeling is given in Winkler (2003, 1993, 1990).

As a closely related case, original microdata **X0** might have been subject to a missing data modeling procedure to obtain completed data **X1a** (Little and Rubin 2002, section 13.4) or edit/imputation modeling procedure to obtain data **X1b** (Winkler 2003) that represent original 'cleaned-up' microdata **X**. The models are created under a *missing-at-random* assumption that is also assumed by hot-deck imputation methods. The basic methods involve imputation only using the EM or various generalizations of EM (Meng and Rubin 1993, Winkler 1993) that have been applied for modeling in the context of statistical matching (D'Orazio, Di Zio, and Scanu 2006) or edit/imputation (Winkler 2003) under either the linear constraints of loglinear modeling or more general convex constraints (Winkler 1990, 1993). New modeling and imputation software (Winkler 2007a) can facilitate the modeling and imputation under a variety of constraints.

The missing data procedures allow us to *disperse* counts associated with 'small' cells to the small cells and the sampling-zero cells. In the simplest situation, we may take the 'observed' count from the set of small cells and disperse it over all the small cells and the sampling-zero cells with straightforward missing data procedures as in Little and Rubin (2002). We can associate given 'observed' counts with a set of small cells with different sets of cells. If we use the convex constraints, then we can place upper and lower bounds on the cells to which we are dispersing the observed counts.

3. Data and Methods

In this section we describe the data of D'Orazio et al. and the modified methods that we use in creating a model **M2** from which to create synthetic data **X2**.

3.1 The Data of D'Orazio et al.

The data of D'Orazio, Di Zio, and Scanu (2006) is a sample of records from a large Italian survey for which only three fields were considered. The fields are AGE, PRO (profession), and EDU (education).

Table 1. Response Categories for Fields

Fields	Transformed Response Categories
Age (AGE)	"0"=15-17 years old; "1"=18-22; "2"=23-64; "4"=65+;
Profession (PRO)	"0"=Manager; "1"=Clerk; "2"=Worker
Education (EDU)	"0"=None or compulsory school; "1"=Vocational school; "2"=Secondary school; "3"=Degree

There are 48 (=4x3x4) data patterns of which many are structural zeros. For instance, a person of age 15-17 cannot have a college degree. The sample size is 2313 that we give as a table of counts and then as a table of probabilities. In Tables 2-10, we use a lexigraphic ordering in which (0,0,0)=0, (0.0,1)=1, …, (3,2,3)=47. We obtain this with the mapping (a1,a2, a3)=a1*12+a2*4+a3*1. The first row of the table is the set of cells 0-7; the second row is the set of cells 8-15, and so on. We use 'z' to represent a structural zero that always has probability zero of being a positive value.

Table 2. Population Counts from Sample File

z	z	z	z	z	z	z	z
15	0	z	z	z	z	z	z
z	1	8	z	27	7	12	z
z	z	142	220	z	123	653	87
759	90	143	2	z	z	4	5
z	0	3	0	12	0	0	0

Table 3. Probabilities for Cells from Sample File

0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.00649	*0.0*	0.0	0.0	0.0	0.0	0.0	0.0
0.0	**0.00043**	0.00346	0.0	0.01167	0.00303	0.00519	0.0
0.0	0.0	0.06139	0.09511	0.0	0.05318	0.28232	0.03761
0.32815	0.03891	0.06182	**0.00086**	0.0	0.0	0.00173	0.00216
0.0	*0.0*	**0.00130**	*0.0*	0.00519	*0.0*	*0.0*	*0.0*

The empirical data are useful due to having only three variables and 48 cells. Because there are 23 structural zeros, we do not have much flexibility in the fitting procedures. With larger, more realistic situations the much smaller proportion of structural zeros will make the fitting much easier. In table 3, the three small (risky) cells (17, 35 and 43) are marked in bold and the sampling-zero cells (9, 41, 43, 45, 46, and 47) are marked in italics.

As a simple alternative that applies conventional loglinear software rather than the more general software used below, we modify cell counts in a manner that gives proportions that are similar to those in the original table, obtain a fitted solution, and can use the fitted solution in generating synthetic data. We generate the synthetic data by drawing records from a table like Table 3 with probability proportional to the size of the probabilities in the table.

3.2 Application of Methods of Edit/Imputation under Convex Constraints

In Table 2, we assume that there are three cells (each having count 3 or less) that we wish to protect from re-identification. A simplistic method of protecting confidentiality might be to take the total of 6 from the three cells and disperse it equally among the three small cells and the six structural zeros. If an intruder cannot decide which cells were small in the original, non-public file, then the intruder might not be able to re-identify.

In larger situations, we might wish to disperse each small cell over several cells in an overlapping manner. To preserve analytic properties, we might use the 'first dispersed' array as the starting point in a general fitting procedure that fits the original interactions from the model for the original, non-public data. The intent is to better preserve analytic properties. The general fitting procedure associates the total of six with the appropriate nine cells in a straightforward missing data fashion as in Little and Rubin (2002, Chapter 13) or more generally Winkler (1993, 2003, 2007a).

If the resultant fitted model still has small cells that correspond to small cells in the original data, then we may wish to use convex constraints to place a priori upper bounds on certain cells. The main issue is whether the fitted model under the convex constraints still preserves the analytic properties quite well. We observe that, if the initial, simple convex constraints do no preserve the analytic properties as well as we would wish, then we have considerable flexibility in applying more sophisticated convex constraints that may better preserve analytic properties (D'Orazio et al. 2006; Winkler 1993, 2007a).

4. Results

In this section, we present results from fitting various loglinear models to the data of section 3. At the first level, it is straightforward to create synthetic data by drawing records from the model. The number of records in the synthetic data need not agree with the number of records in the original data used in creating the model. Because our primary emphasis is on creating synthetic data that allows valid analysis (approximate recreation of the same models and important sufficient statistics), re-identification can still be straightforward because some of the small cells in the synthetic data correspond exactly to small cells in the original, confidential data.

Tables 4 and 5 are obtained by fitting the (starting point) data of Table 3 using 2-way interactions and independence, respectively. By cursory inspection, it is easy to see that the fit of Table 4 is quite close to Table 3 and the fit of Table 5 is quite poor. The approximate Chi-square fits of table 4 are both slightly less than 1 which indicates a good fit (with the caveat that the Chi-square approximation is not always accurate when there are many small cells).

Table 4. Probabilities for Cells after 2-way Interaction Fitting

0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.00649	*0.0*	0.0	0.0	0.0	0.0	0.0	0.0
0.0	**0.00055**	0.00335	0.0	0.01167	0.00291	0.00530	0.0
0.0	0.0	0.06128	0.09523	0.0	0.05306	0.28254	0.03750
0.32815	0.03902	0.06171	**0.00086**	0.0	0.0	0.00184	0.00205
0.0	*0.0*	**0.00118**	*0.00011*	0.00519	*0.0*	*0.0*	*0.0*

Table 5. Probabilities for Cells after Independent Fitting

0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.00611	*0.00038*	0.0	0.0	0.0	0.0	0.0	0.0
0.0	**0.00238**	0.00743	0.0	0.01114	0.00069	0.00215	0.0
0.0	0.0	0.11904	0.03964	0.0	0.07607	0.22046	0.07342
0.33067	0.02044	0.06378	**0.02124**	0.0	0.0	0.00129	0.00043
0.0	*0.00076*	**0.00238**	*0.00079*	0.00358	*0.00022*	*0.00069*	*0.00023*

Using either model (Table 4 or Table 5), we could draw 2313 records with a probability proportion to size. In the first situation, analytic properties would be quite well preserved whereas, in the second situation, the analytic properties of the synthetic data would be very different from the original data. On average, the sampling mechanism preserves the probabilities in the cells. The original data have substantial structural zeros (23 of 48 cells) that restrict the range of the solutions in certain cells. As an instance, cell 8=(0,1,3) has fitted value of 0.0649 that equals the original population value.

By sampling probability proportional to size, we can approximately reproduce the probabilities in individual cells. In particular, cell 39=(3,0,2) when sampled from either Table 4 or Table 5 will typically yield four or less records (0.00129*2313 ≈ 3 0.00184*2313 ≈ 4). If the statistical agency considers cells of size less than 5 a disclosure, then a disclosure occurs whether the synthetic data provide good analytical properties or not.

If the agency decides that all probabilities must be above 0.00173 (corresponding to 4 or more in every cell), then fitting can be done using convex constraints. In particular, we need that P(17)=P(1,1,1)>=0.00173 and P(35)=P(2,2,3)>=0.00173 that are the two small, nonzero cells from the original population. To do this, we need that P(1,1,2)<=0.00216, P(1,2,1)<=00.173, and P(2,1,1)+P(3,1,1)<=0.5145 to assure that cell P(17)>=0.00173. To assure that P(35)>=0.0173, we need P(2,2,1)+P(2,2,2)<=0.09875, P(2,0,3)+P(2,1,3)<=0.13099, and unfortunately P(3,2,3)<= -0.00173. In the original population P(2,2,3)+P(3,2,3)=0.00086+0.00000. Since we must fit to margin P(.,2,3)=0.00086 there is no way to put an upper bound on P(3,2,3) that assures that P(2,2,3)>=0.00173 and P(3,2,3)>=0.00173. If the agency has a rule that all nonzero cells must be above 4, then the most suitable alternative ad hoc solution would be to restrict P(3,2,3)=P(2,2,3)=0.0 rather than restricting both to be greater than 0.00173. To maintain the counts of the total table, 2 (associated with cell (2, 2, 3)) would need to be added to a complementary cell or set of cells.

To create a 'pseudo sample' that is a close approximation of the data of Table 3, we multiply each cell count by 10 and add 4 to all cells. The resultant probabilities are given in Table 6. This has the effect of maintaining most cell probabilities and assuring the 'sampling' zeros are raised to non-zero probabilities.

If we draw a synthetic data set of size 2313 (or 23230 corresponding to the enlarged population of Table 6 or the corresponding fitted Table 7), then, on average, the synthetic data will allow quite accurate reproduction of the analyses that can be performed on the original data of population of Table 3. The advantage of the second procedure is that the user of the data can be specifically informed that a number of the 'small' cells contain artificial counts that are consistent with the analyses that can be performed on the original table. As the user has no way of knowing those cells that have artificial non-zero counts or other cells with small counts that have been slightly increased or decreased, the user has considerably less chance of doing re-identification.

Table 6. Probabilities for Cells after Cell Size Adjustment – 'Pseudo' Population

0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.00663	*0.00017*	0.0	0.0	0.0	0.0	0.0	0.0
0.0	**0.00060**	0.00362	0.0	0.01180	0.00319	0.00534	0.0
0.0	0.0	0.06130	0.09488	0.0	0.05312	0.28127	0.03762
0.32690	0.03892	0.06173	**0.00103**	0.0	0.0	0.00189	0.00232
0.0	*0.00017*	**0.00146**	*0.00017*	0.00534	*0.00017*	*0.00017*	*0.00017*

Table 7. Probabilities after 2-way Fitting to the Values of Table 6

0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.00663	*0.00017*	0.0	0.0	0.0	0.0	0.0	0.0
0.0	**0.00063**	0.00359	0.0	0.01180	0.00316	0.00536	0.0
0.0	0.0	0.06146	0.09472	0.0	0.05308	0.28132	0.03762
0.32690	0.03896	0.06152	**0.00120**	0.0	0.0	0.00174	0.00248
0.0	*0.00019*	**0.00144**	*0.00018*	0.00534	*0.00016*	*0.00035*	*0.00001*

The disadvantage of creating synthetic data from the model of Table 7 is that it may be quite straightforward for an intruder to re-identify by 'guessing' how to reverse the procedures that were used in creating Tables 6 and 7.

To better protect confidentiality, we create Table 8 in which the total 6 from the three small cells (17, 35, and 43) is dispersed among the three small cells and the six sampling- zero (9, 41, 43, 45, 46, and 47) cells. We use Table 8 is the starting point of an EM fitting procedure in which we use standard missing data procedures (Little and Rubin 2002) to fit using the same interactions that we used on the original data of Tables 2 and 3). Table 9 is the limiting solution. Because we may still be able to do re-identification, we reduce the re-identification risk when we fit a model in which upper bounds of 0.0032 in cells 17, 35, and 43. If we draw synthetic data from the fitted solution (Table 10), then re-identification risk is reduced.

Under the dispersal model in which mass 6 is assigned to nine cells, the maximum likelihood is

-1.911699284574. The likelihood with dispersal without convex constraints is = -1.911699284626 and the likelihood with dispersal and convex constraints is -1.911699284627. As these fits are quite good (approximate Chi-square less than 1), any synthetic data produced from the models of Tables 9 or 10 will quite well reproduce analytic properties.

Table 8 Modified Probabilities for Cells from Sample File
Starting point for EM fitting procedure (initial parameters)

0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.00649	*0.00029*	0.0	0.0	0.0	0.0	0.0	0.0
0.0	**0.00029**	0.00346	0.0	0.01167	0.00303	0.00519	0.0
0.0	0.0	0.06139	0.09511	0.0	0.05318	0.28232	0.03761
0.32815	0.03891	0.06182	**0.00029**	0.0	0.0	0.00173	0.00216
0.0	*0.0029*	**0.00029**	*0.00029*	0.00519	*0.00029*	*0.00029*	*0.00029*

Starting points ~0.000288 ((6/9)/2313) for small cells (**bold** - 17, 35, 43) and sampling-zero cells (*italic* – 9, 41, 43, 45, 46, and 47)

Table 9. Probabilities for Cells after 2-way Fitting with dispersal

0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.00649	***0.00015***	0.0	0.0	0.0	0.0	0.0	0.0
0.0	**0.00063**	0.00352	0.0	0.01167	0.00306	0.00512	0.0
0.0	0.0	0.06140	0.09511	0.0	0.05321	0.28225	0.03761
0.32815	0.03887	0.06189	**0.00032**	0.0	0.0	0.00173	0.00216
0.0	*0.00003*	**0.00042**	***0.00005***	0.00519	***0.00018***	***0.00077***	***0.0***

Table 10. Probabilities for Cells after 2-way Fitting with dispersal
Upper Bounds on Cells 17, 35, 43

0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.00649	*0.00028*	0.0	0.0	0.0	0.0	0.0	0.0
0.0	**0.00032**	0.00346	0.0	0.01167	0.00303	0.00519	0.0
0.0	0.0	0.06139	0.09511	0.0	0.05318	0.28232	0.03761
0.32815	0.03891	0.06182	**0.00032**	0.0	0.0	0.00173	0.00216
0.0	*0.00008*	**0.00032**	***0.00009***	0.00519	*0.00030*	*0.00087*	*0.0*

4. Discussion

There is considerable flexibility with the general fitting procedures for larger databases having more variables. Typically, the edit restraints (Winkler 1997, 2003) yield structural zeros that are a much smaller proportion of the number of cells that in the empirical example of this paper. The small proportion makes the dispersal much more straightforward. The dispersal mechanism is quite flexible in that individual small cells can be dispersed over relatively small sets of ‘complementary’ cells. The cells associated with several dispersals can overlap just as they can overlap with general missing data procedures (Little and Rubin 2002, Winkler 2007a).

Although the basic computational procedures are far faster (sometimes several orders of magnitude) than in commercial software (Winkler 2007a), the computational speed is still a major concern in the situations having thirty or more variables. In larger situations, various heuristics and approximations may be appropriate.

If suitable external data are available, then some re-identification should still be possible. If one part of a university released a table such as Table 10 and another part released all (or most) of the two-way tabulations from the original population, then the procedures of this paper (or the more general procedures of Winkler 2007a) could be adapted to create a table using the two sets of constraints that is much closer to the original population (Table 2) than Table 10.

5. Concluding Remarks

This paper provides a procedure for providing a synthetic data set of discrete data that will allow approximate reproduction of analyses from an original, confidential data source. In creating the model for the synthetic data, we use additional convex constraints that are intended to reduce re-identification risk.

1/ This report is 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 author(s) and not necessarily those of the U.S. Census Bureau. The author thanks Philip Steel for a number of comments that led to clarification of several points and the addition of more background information.

References

Agresti, A. (2007), *An Introduction to Categorical Data Analysis (2nd Edition)*, New York, N.Y.: J. Wiley.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006), "Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints," *Journal of Official Statistics*, 22 (1), 137-157.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W., (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.

Little, R. J. A. and Rubin, D. B. (2002), *Statistic Analysis with Missing Data (2nd Edition)*, New York, N.Y.: J. Wiley.

Meng, X.-L., and Rubin, D. B. (1993), "Maximum Likelikehood via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267-78.

Winkler, W. E. (1990), "On Dykstra's Iterative Fitting Procedure," *Annals of Probability*, 18, 1410-1415.

Winkler, W. E. (1993), "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 274-279, also http://www.census.gov/srd/papers/pdf/rr93-12.pdf.

Winkler, W.E. (1997), "Set-Covering and Editing Discrete Data," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 564-569, also http://www.census.gov/srd/papers/pdf/rr9801.pdf.

Winkler, W. E. (2003), "A Contingency Table Model for Imputing Data Satisfying Analytic Constraints," *American Statistical Association, Proc. Survey Research Methods Section*, CD-ROM, also http://www.census.gov/srd/papers/pdf/rrs2003-07.pdf .

Winkler, W. E. (2004), Masking and Re-identification Methods for Public-Use Microdata: Overview and Research Problems, in (J. Domingo-Ferrer and V. Torra, eds.), *Privacy in Statistical Database*, Springer: New York, 231-247, also http://www.census.gov/srd/papers/pdf/rrs2004-06.pdf .

Winkler, W. E. (2005), "Modeling and Quality of Masked Microdata," *American Statistical Association, Proceedings of the Section on Survey Research Method*, CD-ROM, also http://www.census.gov/srd/papers/pdf/rrs2006-01.pdf .

Winkler, W.E. (2007a), "General Methods and Algorithms for Modeling and Imputing Discrete Data under a Variety of Constraints," technical report.

Winkler, W. E. (2007b), "Examples of Easy-to-implement, Widely Used Methods of Masking Data for which Analytic Properties are not Justified," technical report.