RESEARCH REPORT SERIES
(Statistics #2010-02)

General Discrete-data Modeling Methods for Producing Synthetic Data with Reduced Re-identification Risk that Preserve Analytic Properties

William E. Winkler

Statistical Research Division
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: January 28, 2010

General Discrete-data Modeling Methods for Producing Synthetic Data with Reduced Re-identification Risk that Preserve Analytic Properties

William E. Winkler[1], william.e.winkler@census.gov 2008Nov12d
U.S. Census Bureau, Statistical Research Division, Washington, DC 20233-9100

Abstract
General modeling methods for representing and improving the quality of discrete data (Winkler 2003, 2008) extend and connect the editing methods of Fellegi and Holt (1976) and the imputation ideas of Little and Rubin (2002). This paper describes a modeling framework to produce synthetic microdata that better corresponds to external benchmark constraints on certain aggregates (such as margins) and on which certain cell probabilities are bounded both below and above to reduce re-identification risk. Rather than use linear constraints (Meng and Rubin 1993), the modeling methods use convex constraints (Winkler 1990, 1993) in an extended MCECM procedure. Although the produced microdata are not epsilon-private (Dwork 2006, Dwork and Yekhanin 2008), surrogate original microdata would be exceptionally difficult (or impossible) to construct using the standard lp programming procedures of epsilon-privacy.

1. Introduction

This paper describes modeling methods for discrete data. The methods are closely related to general modeling/edit/imputation methods (Winkler 2008) in which models can easily be created using very fast, parameter-driven software. The methods and generalized software are suitable for a wide range of discrete data. The models are used in generalized production edit/imputation software that assure that the 'corrected' data satisfy both edit restraints and preserve joint distributions in a principled manner. Furthermore, the modeling methods use convex constraints (Winkler 1993, 1990) in an EMH algorithm that generalize the linear constraints of the MCECM algorithm of Meng and Rubin (1993). An advantage of the new modeling methods is that the microdata created via the methods can have aggregates that are adjusted to certain benchmark totals.

General convex constraints provide great flexibility in creating models that approximately preserve analytic properties and reduce the re-identification risk in synthetic microdata that are created from the models. Convex constraints allow putting lower and upper bounds on individual cells or on groups of cells. In earlier work, Winkler (2007) showed how to use more elementary methods to reduce re-identification risk by putting lower and upper bounds on both small cells and sampling zeros while still approximately preserving most aggregates needed for loglinear modeling and important joint and conditional probabilities. At that time, Winkler (2007) felt that the risk of re-identification via record linkage experiments was greatly reduced in comparison to data from some previous synthetic-data-generation methods.

Epsilon-privacy represents a gold standard in terms of preventing leakage of information and in preserving privacy. Much research is needed to justify analytic properties of epsilon-private data. Dwork, McSherry, and Talwar (2007b, first two paragraphs of section 5) provide an example from 'census' data in which the amount of noise added to a table having on the order of 1,000,000 cells must be on the order to

1,000,000 (plus or minus) in each cell. In this situation and most others where rigorous epsilon-privacy has been applied, it is not clear that the resultant 'protected' microdata will meet analytic standards acceptable to most economists and statisticians. Additionally, Xaio and Tao (2008) raise serious concerns by demonstrating that it is impractical to verify epsilon-privacy in most situations. Specifically, they prove that L^1-sensitivity of functions (Dwork et al. 2006) is NP-Hard computationally. Dwork et al. (2006) showed that computing the L^1-sensitivity of functions was needed to verify epsilon-privacy in most situations.

The notable exception to the lack of suitable analytic properties is work by Machanavajjhala et al. (2008) that preserves an extended type of epsilon-delta privacy in a very narrowly analytically focused 'on-the-map' application. Machanavajjhala et al. applied clever theoretical techniques and introduced exceptionally complex computational methods that may not be suitable for most general situations.

In this paper, we slightly extend the methods of Winkler (2008, 2007) in a manner that creates a model with a desired set of properties. To do this we place a few pairs of upper and lower bounds on key aggregates needed for the loglinear modeling while placing upper bounds and lower bounds on a very large set of small cells and sampling zeros. The idea is to *target preservation of analytic properties* in the creation of the model. To produce synthetic data, we merely randomly draw from the model in the appropriate fashion. Typically, this means almost exactly preserving the probabilities associated with originally larger cells. Most small cells in the original data are replaced by sets of sampling zeros that have positive probability in the model and that approximately preserve the key aggregates needed for loglinear modeling.

There are several key points of the new methods. First, any direct re-identification experiment will only match originally small cells with sampling zeros that have very small positive probability in the models. Second, because we target preservation of a few analytic properties, we are not creating all of the key aggregates (functions) in a manner where each function satisfies epsilon-privacy. We do create an alternative to a type of epsilon-delta-privacy that we believe would make it exceptionally difficult to reconstruct the original private data in manners suggested by Dwork (2006), Barak et al. (2007), Dwork et al. (2007a) and Dwork and Yekhanin (2008)

Although the computational algorithms needed for creating the models are sufficiently fast for the largest edit/imputation applications, the algorithms need speeding up for even moderate size (50 million cells) modeling situations needed for producing synthetic data.

In the second section of this paper, we give cursory background on edit/imputation and some of the basic computational algorithms. We also describe how a re-identification experiment is performed that assures that private data cannot be easily re-identified but may not satisfy reasonable epsilon-privacy or epsilon-delta privacy. We describe how the models are created. In the third section, we provide empirical results on 'census' data that has be downloaded from the UCI machine learning repository and used in some confidentiality research. Although any synthetic data produced from the model can prevent most re-identification using record linkage and satisfies a condition that can be considered an alternative to very weakened type of epsilon-delta-type of privacy, the synthetic data do not satisfy rigorous epsilon-delta privacy. An interesting experiment (beyond the scope of the present paper) would be for a cryptographer to apply some of the constructive methods (e.g., Dwork 2006, Barak et al. 2007, Dwork et al. (2007a),

Dwork and Yekhanin 2008) to the synthetic data to reconstruct a reasonable approximation of the original private data. The final sections consist of brief discussion and concluding remarks. This experiment would be needed regardless of the type of auxiliary information (Ganta et al. 2008) that might be available to an adversary.

2. Background

In this section we provide background on modeling/edit/imputation, need for computational speed, re-identification using record linkage, and the general iterative fitting algorithm for creating the model.

2.1 Modeling/Edit/Imputation

Modern methods for edit/imputation began with the seminal paper of Fellegi and Holt (1976, hereafter FH). With discrete data, an edit might be that a child of less than 16 could not be married. Their paper provided three principles: (1) The minimum number of fields in each edit-failing record r_0 should be changed to create an edit-passing record r_1 (*error localization*), (2) Imputation rules should be derived automatically from edit rules, and (3) When imputation is necessary, it should maintain marginal and joint distributions of fields.

The FH paper was the first to provide a method that assured that an edit-failing record r_0 could be changed into an edit-passing record r_1. To assure correct error localization, FH showed that implicit edits were needed. Implicit edits are those that can be logically derived from explicitly defined edits. Winkler (1997) provided set-covering algorithms that delineated the implicit edits, the set of which can be considered structural zeros for loglinear modeling. Although a number of statistical agencies have implemented generalized FH production systems that assure the edit-failing records can be 'corrected' to edit-passing records, none have provided FH methods that assure that the records also satisfy joint distributional characteristics from a model. The FH suggestion that hot-deck could be used for (2) and (3) is not possible due to serious deficiencies in hot-deck that were not understood when the FH paper was written (Winkler 2008).

Winkler (2003) provided the theory connecting the edits of FH with the generalized imputation of Little and Rubin (2002). An initial routine (Winkler 1997) finds the set of implicit edits (structural zeros) in a manner that is 100 times as fast as the previous fastest algorithms of Garfinkel, Kunnathur, and Liepins (1986) used by IBM in creating a large system for ISTAT (Barcaroli and Venturi. 1997). A second routine (Winkler 2008) does standard loglinear modeling under a combination of linear and convex constraints in the presence of structural zeros. In the edit setting, the iterative fitting algorithm is a type of EM algorithm as in Little and Rubin (2002). The key aspect of the second routine is having computational algorithms that are sufficiently fast for all of the survey data situations in the statistical agencies. The final routine does the error localization (Winkler 1997) using either branch-and-bound or a greedy algorithm and then fills in missing or 'to-be-changed' values according to the model (contingency table) determined by the second routine. All records are guaranteed to satisfy edits and the overall set of records preserve the probability distributions of the model.

2.2 The EMH algorithm

The general iterative fitting algorithm is extended to an EMH algorithm (Winkler 1993, 1990) for convex constraints that allow putting upper bounds on cells or convex combinations of cells. Because the set of probabilities must add to one, lower bounds can also be put on cell probabilities or simple sums of cell probabilities that might correspond to a marginal constraint. The general EMH algorithm has been used for unsupervised learning of optimal record linkage parameters (Winkler 1993) in which certain probabilities are estimated within restricted ranges based on a priori knowledge. The general EMH algorithm has also been used in statistical matching to create microdata that better corresponds to (external) benchmark constraints (D'Orazio et al. 2006).

In the application of this paper, we apply the EMH algorithm with several constraints. First, we perform standard loglinear modeling to determine the set of interactions needed to get suitably close-fitting model. The model is the final set of probabilities associated with the cells corresponding to the entire set of data patterns. Second, we take the set of counts associated with the small cells (here either 1 or 2) and disperse all of the counts across the entire set of small cells and the entire set of sampling zeros. The intent is to assure positive probability of sampling zeros in a manner that preserves most of the characteristics of the best-fitting set of interactions under purely linear constraints. Third, we place upper bounds (say 0.000004) on the probabilities associated with the originally small cells that assure that the final fitted probabilities are zero to five decimal places. Fourth, if necessary, we can place upper and lower bounds on a few of the marginal probabilities in the final fitted contingency table that deviate substantially from the marginal probabilities in the original, private data.

To create the synthetic data, we randomly draw from the contingency table probability proportional to size. If necessary, we can create multiple copies of the synthetic data.

2.3 Re-identification via Record Linkage

After modeling and creation of synthetic data **Y** from the original data **X**, we can perform re-identification experiments. To do this we merely match data **Y** directly against data **X**. The re-identification experiment is conservative in the sense that any intruder would likely have data **Y1** that is more difficult to match against **X** than **Y**. In a real-world situation, the intruder would have names and other identifying information associated with individual records in data **Y1**. Based on the worst-case re-identifications, it is possible to extrapolate downward explicit re-identifications of individual records or of overall re-identification rates. The downward extrapolation can be based on assumed typographical error rates or the record linkage metrics that are used to compare individual fields. With discrete data, we might only do exact comparison of individual fields and use an EM-latent class algorithm for estimating the best record linkage parameters. Kim and Winkler (1995) and Winkler (1998) used the EM algorithm and different field-comparison metrics for re-identification with continuous data. For convenience, we assume that we are using entire populations so that we need not extrapolate for different sampling scenarios. If we use an entire population, matching is much easier and an upper bound on re-identification risk is more easily computed.

Any record corresponding to a small cell in the data **Y** that can be associated via record linkage with the correct corresponding cell in **X** with high matching probability can be considered a re-identification. With continuous data scenarios, both Fuller (1993) and Winkler (1998) showed how to perform the matching to get explicit re-identification.

Discrete-data re-identification is much more straightforward under the complete population scenario of this paper. Typically, if we randomly draw synthetic data from the model of section 2.2, *we will not get any re-identification* using record linkage. The key issue with the synthetic data is whether the synthetic data preserves a few analytic constraints so that someone using the synthetic data **Y** would approximately reproduce results that could be obtained from the original data **X**.

With epsilon-privacy (e.g., Dwork 2006), individuals make similar assumptions about the best possible data **Y1** (or **Y**) that might be matched against data **X**. Epsilon-privacy goes further in that it assures almost no leakage of information that prevents re-identification but does not presently preserve analytic properties in any clearly established manner. Ganta et al. (2008) explicitly bring in the use of auxiliary information in demonstrating that epsilon-privacy prevents any type of re-identification.

2.4 The Empirical Data and Restraints Used for Modeling

Data are from the University of California at Irvine machine learning respository. The specific data set is 'Adult'. The variables (fields) downloaded were age, WorkClass (8 values), Education (16 values), MaritalStatus (7 values), Occupation (14 values), Race (5 values), Sex (2 values), and Country (41 values). For initial testing purposes, we used WorkClass (7 values), MaritalStatus (7 values), Race (5 values), and Sex (2 values) that yielded 490 ($7 \times 7 \times 5 \times 2$) data patterns. There are 45221 data records and there are no missing fields within data records. WorkClass is reduced to 7 values because one of its values (NoWork) never occurs in the data set.

The data have 80 small cells having count 1 or 2, 191 cells that are sampling zeros, and 290 cells having count above 2. The total count associated with the small cells is 103. We determine that the all 3-way interaction model gives good fits with linear constraints only. We use an EM fitting procedure in which we disperse the total count of 103 associated with the small cells across all 271 (80 + 191) cells having small or zero counts. The starting value is 103/271 in each cell and the expected E-values are based on the current set of the parameters from the M-step. The counts of the larger cells are not varied in the modeling because we are assuming that we will not be able to effectively re-identify individual large cells in synthetic data **Y** randomly drawn from the model with the individual large cells in data **X**. After the initial fitting under linear restraints, we repeat the fitting where we place additional convex constraints (upper bounds of 0.000004) on the small cells. The synthetic data is created reproducing the counts of the non-small cells and randomly sampling from the remaining cells (both small and sampling zeros) with a probability proportional to size procedure until we achieve synthetic data **Y** of size 45221.

In earlier work, Winkler (2007) showed that the fitting and modeling methods had great flexibility in a small situation representing 48 ($4 \times 3 \times 4$) cells where nearly half of the cells were structural zeros. In more recent work, Winkler (2008) showed that the modeling methods had somewhat greater flexibility in a situation with 96 ($4 \times 3 \times 4 \times 2$) cells. The point is that, with the smallest situations, we have very little flexibility in the modeling to preserve the analytic properties. With more cells (490 data patterns), we have considerably greater flexibility in preserving analytic properties. With an even greater number of cells (580,160 = $74 \times 7 \times 7 \times 16 \times 5 \times 2$), we have even greater

flexibility in preserving analytic properties but may encounter computational issues (10 minutes for the general fitting procedure to converge).

3. Results

The results presented in this section are intended to represent a small situation (490 cells or data patterns) that is still quite cumbersome to present because of the large size of the tables. We present the 490-cell situation because we believe that it is adequate for illustrating how analytic properties are preserved while significantly reducing re-identification risk.

Fitting the 3-way interaction model **M1** (with linear but no convex constraints), we have that the maximum possible likelihood is -3.234682 and that the likelihood that we achieve is -3.234982. The maximum deviation allowed by the fitting software is 0.0000000000100. If we fit with the same interaction restraints and an additional restraint with an upper bound of 0.000004 on each originally small cell (model **M2**), we get the likelihood of -3.241030 that indicates a reasonably good overall fit. As our fitting uses all 3-way interactions, we need to examine how closely the 3-way margins from the limiting solution under model M2 agree with the 3-way margins from the original data. In indexing cells, we use a lexicographic ordering in which (0,0,0,0)=0, (0.0,0,1)=1, …, (6,6,4,1)=489. We obtain this with the mapping (a1, a2, a3, 4)=a1*24+a2*8+a3*2+a4*1. If $\mathbf{X}_i$, $1 \le i \le 4$, is the i^{th} variable, then $\{X_1=i_1, X_2=i_2, X_3=i_3, X_4=i_4\} = (i_1, i_2, i_3, i_4)$.

Table 1 represents original and fitted probabilities associated with a few selected individual cells. It is an excerpt from the full Table A.1 given in the Appendix. A cell with a count of 1 has probability 0.00002 and a cell with count of 2 has probability 0.00004. All of the probabilities in the table are rounded to five digits. Cells 0000-0007 show that the individual cell probabilities are reasonably close to each other. Cells 0020, 0021, and 0301 have the largest deviations. Cell 0107 is an original cell with count 1 that is given a fitted probability above zero and below 0.000004. Cells 0485-0489 are sampling zeros that are given a positive probability of approximately 0.00001. When we randomly sample from Table A.1, we have positive probability of sampling each cell but originally small cells will seldom appear in the set of synthetic records. All of the greatest deviations are associated with cells that have total probability of less than 0.003. The greatest multiplicative deviation in the remaining cells is well less than 1.0. The key issue is how well the margins are preserved.

Table 1. Original and Fitted Probabilities for Selected Cells

Cell	Original	Fitted
0000 0 0 0 0	0.02859	0.02876
0001 0 0 0 1	0.25344	0.25328
0002 0 0 1 0	0.00172	0.00163
0003 0 0 1 1	0.00781	0.00790
0004 0 0 2 0	0.00031	0.00037
0005 0 0 2 1	0.00181	0.00175
0006 0 0 3 0	0.00042	0.00042
0007 0 0 3 1	0.00210	0.00210
0020 0 2 0 0	0.09670	0.09636

0021 0 2 0 1	0.12426	0.12460
0107 1 3 3 1	0.00002	0.00000
0301 4 2 0 1	0.00637	0.00610
0485 6 6 2 1	0.00000	0.00001
0486 6 6 3 0	0.00000	0.00001
0487 6 6 3 1	0.00000	0.00001
0488 6 6 4 0	0.00000	0.00001
0489 6 6 4 1	0.00000	0.00001

Table 2 contains a few selected marginal probabilities for variables 1, 3, and 4. The largest deviations 0.000210, 0.00105, and 0.000100 occurred at marginal cells 0067, 0014, and 0054, respectively. No other specific marginal probabilities for the other interaction patterns were this large. We also give the first eight marginal probabilities. Examination of table A.2 indicates that most marginal probabilities from the fitted data are very close to the marginal probabilities from the original data. The closeness of the marginal probabilities indicates that association-rule mining and other elementary analyses of the joint and conditional probabilities should yield results from synthetic data created from Table A.1 that agree somewhat with comparable results from the original confidential data.

Table 2. Original and Fitted 3-way Margins
for Selected Marginal Cells

Pattern = 3, Variables 1,3,4		
00000	0.205988	0.205988
00001	0.427102	0.427102
00002	0.007607	0.007589
00003	0.013511	0.013518
00004	0.002211	0.002223
00005	0.003936	0.003925
00006	0.002410	0.002423
00007	0.004179	0.004146
00014	0.000133	0.000028
00054	0.000199	0.000099
00067	0.000000	0.000210

4. Discussion

Re-identification experiments may not be effective in proving the privacy of synthetic data produced according to the methods of this paper. The synthetic data do not appear to satisfy any rigorous type of epsilon- or epsilon-delta privacy. If a cryptographer were to reconstruct a moderate subset of the originally-private microdata from the synthetic data, then the reconstruction should prove that re-identification experiments are not valid in verifying the privacy of synthetic microdata in most situations.

Any reconstruction of the original data from the synthetic data would be computationally challenging in moderate size situations. In the 6-variable scenario, there are 588,160 data patterns, 9447 cells having counts of 1 or 2, and 3098 cells having counts of greater than 2. The total from all the cells is 45221. Because there are so many

sampling zeros (~98% of 588,160 possible cells), we have great flexibility in assigning positive probabilities to the sampling zero cells in a manner in which analytic properties are approximately preserved (much better than with the 490-cell example of this paper). After the random sampling, we have a synthetic data set (or multiple synthetic data sets) in which the small counts from 9447 cells in the original private data are placed in a suitable set of sampling zero cells.

More research needs to be done on what it means to preserve analytic properties. In particular, there needs to be more agreement among researchers on what it means to preserve analytic properties. This paper merely shows that the overall fit of the data and almost all of the 3-way margins having larger probability agree quite closely between the fitted and original data.

The computational algorithms need to be speeded up and altered. In testing on the larger data (588,160 cells), the fitting with both linear and a very simplified set of convex constraints needed 10 minutes CPU time. With a very large set of convex constraints and a variant of the current set of algorithms for the convex constraints, the fitting takes 10-100 times as long.

5. Concluding Remarks

This paper provides methods for modeling discrete data that generalize standard loglinear modeling to methods that also include convex constraints. When properly applied, the convex constraints allow significantly reduced chance of re-identification using record linkage methods. The synthetic data randomly drawn from the models approximately (but very closely) preserve a few analytic characteristics whereas epsilon-privacy methods (Dwork et al. 2007b, first two paragraphs of section 5) have not been demonstrated to preserve analytic properties. The synthetic data created by the methods of this paper do not necessarily satisfy epsilon-privacy or epsilon-delta-privacy (Machanavajjhala et al. 2008) but might be exceptionally difficult to re-identify using cryptographic protocols and exceptionally large amounts of computation.

References

Abowd, J. M. (2008), "How Protective are Synthetic Data," *Privacy in Statistical Databases 2008*, New York: Springer.

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

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., and Talwar, K. (2007), "Privacy, Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release," PODS '07, Beijing, China.

Barcaroli, G., and Venturi, M. (1997), "DAISY (Design, Analysis and Imputation System): Structure, Methodology, and First Applications," in (J. Kovar and L. Granquist, eds.) *Statistical Data Editing, Volume II*, U.N. Economic Commission for Europe, 40-51.

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

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.

Dwork, C. (2006), "Differential Privacy," 33rd International Colloquium on Automata, Languages and Programming – ICALP 2006, Part II, 1-12.

Dwork, C. (2008), "Differential Privacy: A Survey of Results," in (M. Agrawal et al., eds.) TAMC 2008, LNCS 4978, 1-19.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006), "Calibrating Noise to Sensitivity in Private Data Analysis," 3rd Conference on Cryptography – TCC 2006, 365-384.

Dwork, C., McSherry, F., and Talwar, K. (2007a), "The Price of Privacy and the Limits of LP Decoding," STOC '07, San Diego, CA.

Dwork, C., McSherry, F., and Talwar, K. (2007b), “Differentially Private Marginals Release with Mutual Consistency and Error Independent Sample Size,” UNECE Worksession on Statistical Data Confidentiality, Manchester, UK,at http://www.unece.org/stats/documents/2007/12/confidentiality/wp.19.e.pdf .

Dwork, C. and Yekhanin, S. (2008), “New Efficient Attacks on Statistical Disclosure Control Mechanisms,” Advances in Cryptology—CRYPTO 2008, to appear, also at http://research.microsoft.com/research/sv/DatabasePrivacy/dy08.pdf .

Ganta, S., Prasad, S., and Smith, A. (2008), Compositional Attacks and Auxiliary Information in Data Privacy,” ACM KDD ’08, 265-273.

Fellegi, I. P., and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association*, 71, 17-35.

Fuller, W. A. (1993), “Masking Procedures for Microdata Disclosure Limitation,” *Journal of Official Statistics*, 9**,** 383-406 (http://www.jos.nu/Articles/abstract.asp?article=92383).

Kim, J. J., and Winkler, W. E. (1995), “Masking Microdata Files,” *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 114-119 (http://www.amstat.org/sections/SRMS/Proceedings/papers/1995_017.pdf , longer report http://www.census.gov/srd/papers/pdf/rr97-3.pdf) .

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

Liu, C. (2000), “Estimation of Discrete Distributions with a Class of Simplex Constraints,” *Journal of the American Statistical Association*, 95 (449), 109-120.

Machanavajjhala, A., Kifer, D., Abowd, J. Gehrke, J., and Vilhuber, L. (2008), “Privacy: Theory meets Practice on the Map,” ICDE 2008, 277-286.

Meng, X.-L., and Rubin, D. B. (1993), “Maximum Likelihood 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," P*roceedings 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. (1998), “Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata,” *Research in Official Statistics*, 1, 87-104, http://www.census.gov/srd/papers/pdf/rrs2005-09.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. (2007), “Analytically Valid Discrete Microdata and Re-identification,” available at http://www.census.gov/srd/papers/pdf/rrs2007-19.pdf .

Winkler, W. E. (2008), “General Methods and Algorithms for Modeling and Imputing Discrete Data under a Variety of Constraints,” Statistical Research Division Report RRS2008-08, available at http://www.census.gov/srd/papers/pdf/rrs2008-08.pdf .

Xiao, X., and Tao, Y. (2008), “Output Perturbation with Query Relaxation,” VLDB, 857-869.

Longer technical report with full appendix is available at http://doku.iab.de/fdz/events/2008/SDC-Workshop_Winkler_Paper2.pdf .