WP No. 26
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (ii): Developments related to new methods and techniques

A Contingency-Table Model for Imputing Data Satisfying Analytic Constraints

Supporting Paper

Submitted by [Bureau of the Census, USA][1]

ABSTRACT

This paper describes a method for imputation in general contingency tables when the imputations are subject to both analytic (edit) constraints and probabilistic distributional constraints. The model extends edit ideas in Fellegi and Holt (1976) and Winkler and Chen (2002). The model extends missing-at-random imputation ideas in Little and Rubin (1987). Some of the ideas are related to Friedman (2001) and Thibaudeau and Winkler (2002).

Keywords: hot-deck, loglinear models, set-covering

I. INTRODUCTION

1. Missing and contradictory data are endemic in computer databases. If we are only interested in missing data, then the methods described in Little and Rubin (1987) represent a good starting point for suitable imputation. In this paper, we typically consider imputation under a missing at random (MAR) mechanism for nonresponse. As we illustrate later, in the simplest situations, filling in for nonresponse under MAR corresponds to the usual hot-deck methods for filling in missing data provided an exceptionally large number of donors are available. If an exceptionally large number of donors is not available, then we must use model-based (model-assisted) methods as in Little and Rubin

2. If we are only concerned with contradictory information, then we might begin by editing the data. In editing, we are concerned with "correcting" data in records according to the rules defined by analysts. For instance, we might not allow an individual of less than 15 years of age to be married. To correct a record subject to this error condition, we might change age to be greater than or equal than 15 or change marital status to not married. Our edit ideas build on the edit model introduced by Fellegi and Holt (1976). Although the original model appeared in the

[1] Prepared by [william.e.winkler@census.gov 2003Aug24]

statistical literature, the methods for implementing it have primarily relied on methods from the Operations Research (OR) literature and are largely unknown to statisticians.

3. The goal of this paper is to provide methods for filling in for missing data and replacing contradictory data in a manner than is consistent with the missing-data methods of Little and Rubin (1987) and the edit methods of Fellegi and Holt (2001). The outline of this paper is as follows. In the second section, we give background on hot-deck, the missing-data methods as given in chapter 9 of Little and Rubin (1987), and the Fellegi-Holt model of editing. The hot-deck method provides crucial insights in the MAR imputation used by Little and Rubin. When the hot-deck is properly structured, it can also provide insight into the easiest aspects of the edit models. A key feature of the global optimization methods of the Fellegi-Holt model is that it gives a method of filling in missing values or replacing contradictory values in a manner that assures the resultant "corrected" record satisfies all edits. By adapting the modeling ideas of Little and Rubin, we also provide a method that places a probability distribution on the set of "corrected" records that is consistent with the observed complete and incomplete data. In the third section, we present the main theoretical results. Most of the ideas are straightforward but have not been connected previously. Section four consists of discussion and the final section is concluding remarks.

II. **BACKGROUND**

4. This section provides background on the hot-deck, the MAR missing-data mechanism, and the Fellegi-Holt model of editing. In this paper, we only address situations with discrete data. Extensions to continuous or combinations of discrete and continuous are sometimes straightforward.

II.1 Hot-Deck

5. Let $\boldsymbol{X}$ be an $n \times k$ array of discrete data. For some $m < n$, the first m rows of array X are complete data. The remaining n-m rows of $\boldsymbol{X}$ are incomplete data in which some of the data fields are missing. We use $\boldsymbol{X}_j$ to denote the j^{th} row that corresponds to a data record. We can think of $\boldsymbol{X}$ as being obtained from a sample survey or census. The missing data are due to item nonresponse. For $j > m$, $\boldsymbol{X}_j = (x_{j1}, \ldots, x_{jk})$ has some x_{ji} values missing. We denote the missing values of x_{ji} in row j by NR(j) and the non-missing values by R(j).

6. For the $\boldsymbol{X}_j$, $j>m$, we match against the corresponding values of x_{ki} in R(j) to the records X_{jl} for $l \leq m$. Denote the set of records that match X_j according to this criteria by M(j). Then M(j) can contain no record, one record, or more than one records. In the first case, hierarchical collapsing rules for using subsets of R(j) in the matching are used. If collapsing has been used, then we denote M(j) by $M_c(j)$. If only one record is in M(j), then in many practical situations, the same record will be in $M(j_1)$ for some $j_1 > m$ and $j_1 \neq j$. This is the well-known situation where a complete-data record serves as a donor for more than one incomplete record. If there is more than one record in M(j), then it is still typical that the filled in values may not accurately represent the distributions of the values of the variables from the non-missing (complete) data records. This can be because the inherent sampling mechanism induced by matching of the non-missing values of record $\boldsymbol{X}_j$ may not obtain a set of records that accurately represent the

distributions of the values for the missing data items. We use $\boldsymbol{X}_{CH}$ to denote an array where the missing item values have been filled in by hot-deck.

7. In many practical situations, hot-deck will yield marginal distributions for single variables in the completed data $\boldsymbol{X}_{CH}$ that approximate the marginal distributions from the original complete data from the first *m* rows of $\boldsymbol{X}$. It is well known that joint marginal distributions are not preserved. This is often due to the collapsing rules that are used in the matching. If there is much collapsing (M(j) null for many *j*) or many single donor situations, then even the marginal distributions for single variables can be somewhat distorted. We observe that the matching on the observed (non-missing) values of variables is a type of sampling mechanism that may not typically get values of variables that represent reasonable approximations of the true underlying set of values of the variables.

8. In the next section, we will apply missing data ideas that are consistent with hot-deck and give us a framework for better understanding of the characteristics of the joint distributions that we need.

II.2 MAR Missing-Data Mechanism and Imputation

9. To focus our understanding better, we can let $\boldsymbol{X}_T$ represent a large population in which there is no missing data. We let $f_T(x_1, \ldots, x_k)$ represent the probability distribution of the data. We let $\boldsymbol{X}$ represent our observed population data in which the rows greater than *m* have missing item values. We wish to fill-in data in $\boldsymbol{X}$ to get a completed set of data $\boldsymbol{X}_C$ with no missing data and for which the associated probability distribution $f_C(x1, \ldots, x_k)$ is a good approximation of $f_T(x_1, \ldots, x_k)$. For $j > m$, let $\boldsymbol{X}_j = (x_{j1}, \ldots, x_{jk})$ be a record with missing items. Let M(j) be the set of records among the complete data records that it can be matched against. For now, we assume that M(j) contains a moderately large number of records. If we sample one record in M(j) at random, then we fill in values in X_j assuming that the item value are missing at random. The missing values depend on the non-missing values in $\boldsymbol{X}_j$ and not on the (unobserved) missing values in $\boldsymbol{X}_j$.

10. We next raise a somewhat subtle point that allows us to sample in a manner that also causes the filled-in data $\boldsymbol{X}_C$ with distribution f_C to satisfy edit constraints. We begin with the hypothetical situation where there are a nearly infinite number of donors for each record that requires imputation. If all the donor records are only taken from records that satisfy edits, then the record that results from the substitution of item values from the donor record will also satisfy edits. Indeed, the resultant record will be identical to the donor record. In practical situations, we must do collapsing to get potential donors. With many real-world situations, we will not be able to find suitable donors for greater than 99.9% of the records. We return to this issue after we describe MAR imputation.

11. We fill in missing data via an EM procedure that ignores edit constraints. The procedures are those due to Little and Rubin (1987, section 9.4). There are two ways of doing this. The first way is to find a set of interactions that parsimoniously represents the data. We can use only the first *m* rows to model the data as in Bishop, Fienberg, and Holland (1975). With the set of interactions found in the first modeling exercise, we can also fill-in the missing data in the last *n*-

m rows as in Little and Rubin (1987, Chapter 9). In this latter situation, we can also increase the number of interactions based on terms occurring in the last *n-m* rows. We also adjust the margins to which the fitting is done based on the additional non-missing items in the last *n-m* rows. The result of the two modeling steps is a representation $\mathbf{X}_C$ of data $\boldsymbol{X}$ in which missing values are replaced by expected values that may not be integers.

12. We restate the above filling-in procedure as a lemma.

Lemma 1 (Little and Rubin 1987). Let $\boldsymbol{X}$ be an incomplete data array. Let I_1 be a set of interactions and let M_1 be a set of margins determined by the complete data records and the non-missing items in the in the incomplete records of $\boldsymbol{X}$. Then the EM fitting procedure yields a complete data representation $\boldsymbol{X}_C$ that is a model for filling-in incomplete data records with probability distributions that are consistent with the observed data $\boldsymbol{X}$.

13. There are several observations that we can now make. First, the resultant representation $\boldsymbol{X}_C$ along with appropriate interaction parameters and margins that were used in the fitting is a parametric form (i.e., model for the data). For convenience, we assume that $\boldsymbol{X}_C$ is a probability distribution by dividing each cell value by an appropriate population value that causes the resultant set of cells to add to one. Second, the number of cells in the array is given by $S_{\boldsymbol{X}} = |v_1| |v_2| \cdots |v_k|$ where $|v_i|$ is the cardinality of the set of values for field *i* in the observed data. The number S_X can be very large (approaching 10^{46}) for a large labor force survey (Winkler 1997). For each record $\boldsymbol{X}_j$, $j > m$, that has missing item values, we can enumerate cells $M_{\boldsymbol{X}}(j)$ that match on the nonmissing values of $\boldsymbol{X}_j$. We note that, unlike the hot-deck situation, it is always possible to find donors satisfying all possible missing data patterns. If we randomly draw a value from $M_{\boldsymbol{X}}(j)$ with probability proportional to the probability of the cell in the fitted array $\boldsymbol{X}_C$, then we are preserving the probability distribution defined by the probability distribution of $\boldsymbol{X}_C$. We note that the sampling mechanism in $M_{\boldsymbol{X}}(j)$ causes the conditional probabilities to add to 1.

14. The difference between this method and hot-deck is that hot-deck often has 0 or 1 donors instead of potentially thousands or more with this procedure. The collapsing of typical hot-deck implementations can cause distortions in the joint distributions. At the expense of significantly increased computation in contrast to typical hot-deck situations, the data structure given by $\boldsymbol{X}_C$ allows us the possibility of filling in data in a manner that satisfies restraints on the probabilistic distributions and additional "edit" restraints.

II.3 Editing Model of Fellegi and Holt

15. Fellegi and Holt (1976, hereafter FH) provided a theoretical model for editing. In providing their model, they had three goals:

1. The data in each record should be made to satisfy all edits by changing the fewest possible variables (fields).
2. Imputation rules should derive automatically from edit rules.
3. When imputation is necessary, it should maintain the joint distribution of variables.

16. Fellegi and Holt (Theorem 1) proved that implicit edits are needed for solving the problem of goal 1. Implicit edits are those that can be logically derived from explicitly defined edits. Implicit edits contain information about edits that do not fail initially for a record but may fail as values in fields associated with failing edits are changed. An edit places restrictions on certain fields called *entering* fields. By "correcting" a record, we mean changing values in entering fields associated with failing edits so that the modified record no longer fails edits. Goal 1 is referred as the *error localization* (*EL*) problem. The *complete* set E of edits consists of the set of explicit and implicit edits.

17. Prior to the seminal work of Fellegi and Holt (1976), edit methods would fail to correct many records. They failed because, as values associated with fields in explicit edits were changed, previously non-failing edits would fail. Both the set-covering algorithms for generating implicit edits and the integer programming methods of finding optimal solutions are known to be NP-Complete (Garfinkel, Kunnathur, and Liepins 1986). Much of the later work (Winkler 1995, 1997; Chen 1998; Winkler and Chen 2002) has been concerned with increasing the speed of the algorithms. In the extreme case of a large labor force survey, the computational speedups can be on the order of 100,000. In those situations, the new methods are sufficiently fast for production editing systems that fill-in data that satisfy edit restraints. Prior methods, however, do not yield filled-in data that satisfy probabilistic restraints assuring the joint and marginal distributions correspond to the originally observed data.

18. The following example illustrates some of the computational issues. An edit can be considered as a set of points. Let edit E = {married & age $\leqq$ 15}. Let *r* be a data record. Then *r* $\in$ E => *r* fails edit. This formulation is equivalent to 'If age $\leq$ 15, then not married.' We note that if a record *r* fails a set of edits, then one field in each of the failing edits must be changed. Now consider an implicit edit E_3 that can be implied from two explicitly defined edits E_1 and E_2; i.e., E_1 & E_2 => E_3.

E_1 = {age $\leq$ 15, married, . }
E_2 = {. ,not married, spouse}
E_3 = {age $\leq$ 15, . , spouse}

19. In the above edits, the values in the entering fields age, marital status and relationship to head of household are restricted. Implicit edit E_3 can be logically derived from E_1 and E_2. If E_3 fails for a record *r* = {age $\leq$ 15, not married, spouse}, then necessarily either E_1 or E_2 fail. Assume that the implicit edit E_3 is unobserved. If edit E_2 fails for record *r*, then we may change the marital status field in record *r* to married to obtain a new record r_1. Record r_1 does not fail for E_2 but now fails for E_1. If the implicit edit E_3 were observed, then we would know to change at least one additional field in record *r*. For much larger data situations having more edits and more fields, the number of possibilities increases at a very high exponential rate.

20. The main theorem of Fellegi and Holt proved that any cover C_1 of the fields in the failing (explicit and implicit) edits associated always yields an edit-passing record r_1 from record *r* by finding new values of the fields in C_1. If the cover C_1 is prime (i.e., has no subsets that are also covers), then we know that we must always change the value in each field in the cover. Efficient

algorithms for filling in data are available in Chen (1998) and Chen, Thibaudeau, and Winkler (2002).

21. We are primarily concerned with enhancing the Fellegi-Holt in the sense of better preserving probabilistic restrictions of the characteristics on the joint distributions of the filled-in data represented by $\boldsymbol{X}_C$. If a value in a field is designated for replacement because of the edit restraints, then we set it to missing. This artificial missing condition (called *replaced* or *blanked*) and the ordinary missing data must be imputed. The restriction on the imputation is that the result of the imputation in a record must not create a record that fails edits. The difficulty with conventional edit/imputation built around a Fellegi-Holt edit mechanism is that formal models for the imputation process have not been available. If hot-deck methods are combined in an ad hoc way with the edits, then the resultant data does not satisfy distributional constraints and many imputed records may still fail edits. In practice, the result of imputations in the Fellegi-Holt systems has been data that do not satisfy goal 2 of Fellegi and Holt and often do not satisfy goal 3. Satisfying goal 3 might cause additional constraints to be placed on the joint distributions. With the types of OR approaches in applications of the Fellegi-Holt model that have typically been used, it is not clear how these types of constraints could be imposed.

III. **THEORETICAL RESULTS**

22. This section contains the main theoretical results. For clarity and to target a few additional ideas specifically, we repeat some of the notation from previous sections. Let $\boldsymbol{X}$ be an $n \times k$ array of discrete data. For some $m < n$, the first m rows of array $\boldsymbol{X}$ are complete data. The remaining n-m rows of $\boldsymbol{X}$ are incomplete data in which some data are missing. We can think of $\boldsymbol{X}$ as being obtained from a sample survey or census. Some of the missing data is due to item nonresponse. The remainder of the missing data is due to fields that have been blanked because of edit constraints. Edit constraints can be given as in the Fellegi-Holt model of statistical data editing (Fellegi and Holt 1976). We let $\boldsymbol{X}_C$ represent the filled-in or completed data that will be obtained according to procedures that we describe below.

III.1 Current Means of Imputation in the Fellegi-Holt Model of Editing

23 If all implicit edits are available, then it is straightforward to fill-in the missing or to-be-replaced values of a record (see e.g., Winkler and Chen 2002) in an OR sense. The filling-in procedure may not necessarily preserve statistical distributions. For any set of missing (or replaced) items in a record $\boldsymbol{X}_j$ for $j > m$, we can fill in the missing values sequentially according to method 1 of Fellegi and Holt (1976). If we repeat the fill-in procedure, then it is possible to enumerate all the ways the record $\boldsymbol{X}_j$ can be filled in to yield completed records $\boldsymbol{X}_j^{c(i)}$ where $0 \leq i \leq N(j)$ where $N(j)$ is the number of different ways that X_j can be filled-in. The Fellegi-Holt methodology does not provide a suitable means of putting probability distributions on the set of filled-in values. From Winkler and Chen (2002), we know that there is potentially a very large combinatorial explosion of the number of values that can be filled-in. Further, we can observe that there is little possibility that the set of hot-deck donors will give a full representation of the filled-in values in $\{\boldsymbol{X}_j^{c(i)}$ where $0 \leq i \leq N(j)\}$ or that the hot-deck donors will even have a sufficiently large number of different value states to preserve a large set of explicitly defined margins.

III.2 Missing Data Imputation that Satisfies Edit Restraints

24. It is straightforward to extend Lemma 1 to the situation in which we include both restraints due to edits and to probability distributions. We assume that we are able to generate all the needed edits by separate procedures and that the set of edits is consistent. We assume that each of the edits define a structural zero (Bishop, Fienberg, and Holland 1975) of the fitting procedure.

25. An edit restraint imposes a structural zero because certain combinations of values of fields corresponding to the edit are forbidden. The appropriate margin and all corresponding cells in the $\boldsymbol{X}_C$ must be zero. Structural zeros can be dealt with according to the procedure in Lemma 1 of Winkler (1990). If the set of marginal restraints and the set of structural zeros induced by the edits yield a logically consistent set of constraints, then the iterative fitting procedure will still converge in the sense the likelihood will increase monotonely. Winkler (1993) provides a more general iterative fitting procedure than the special case needed to yield the following lemma.

Lemma 2. Let $\boldsymbol{X}$ be an incomplete data array. Let I_1 be a set of interactions, let M_1 be a set of margins determined by the complete data records and the non-missing items in the in the incomplete records of $\boldsymbol{X}$ and let E_1 be a complete set of edit restraints. Then the EM fitting procedure that accounts for E_1 as structural zeros yields a complete data representation $\mathbf{X}_C$ that is a probabilistic model for the data.

26. We observe that the only difference between the iterative fitting procedures of Lemma 1 and 2 is due to the method of accounting for structural zeros. Although $\boldsymbol{X}_C$ obtained by Lemma 1 and $\boldsymbol{X}_C$ obtained by Lemma 2 represent all S_X potential cells in the product space associated with the fields of X, the latter mechanism forces a large number of cells to zero and automatically adjusts the remaining cell probabilities associated with $\boldsymbol{X}_C$.

27. Our goal in the remainder of this paper is to develop a imputation strategy that:
(1) imputes integer values of missing items,
(2) preserves the probability structure of the model that produced $\boldsymbol{X}_C$, and
(3) imputes records that satisfy edits.

The EM procedure for filling in missing data assumes that a set of interactions between variables has been determined. Little and Rubin (1987, Chapter 9) provide good examples on how the interactions and the EM fill-in procedure are done.

28. Given the representation X_C, we can match any incomplete record $\boldsymbol{X}_j$ against the appropriate rows in $\boldsymbol{X}_C$. We denote the matching rows by $M(j) = \{X_{C,1}, X_{C,2}, \ldots, X_{C,N(j)}\}$. If we then sample from M(j) with probabilities proportional to the probabilities in the appropriate rows in $\boldsymbol{X}_C$, then we obtain records that satisfy edits and preserve the probabilistic structure induced by $\boldsymbol{X}_C$. This yields the following theorem.

Theorem 1. Let $\boldsymbol{X}$ be an incomplete data array. Let I_1 be a set of interactions, let M_1 be a set of margins determined by the complete data records and the non-missing items in the in the incomplete records of $\boldsymbol{X}$, and let E_1 be a complete set of edit restraints. Then, using a complete

data representation $\boldsymbol{X}_C$ obtained via the EM fitting procedure of Lemma 2, it is possible to fill in incomplete data records in a manner that preserves the probability distribution of $\mathbf{X}_C$ and that satisfies edit restraints.

29. *Remark.* The procedure is not intended to be computationally feasible in the very largest situations. It is intended to cast theoretical insight on an imputation method for contingency tables that satisfies edit constraints and preserves the underlying probability distribution. Although Chen et al. (2003) provide computationally tractable methods for enumerating all of the different ways of filling in a record, the iterative proportional fitting methods for getting the complete data representation $\boldsymbol{X}_C$ are not computationally tractable when $\boldsymbol{X}_C$ has on the order of 10^{35} cells.

IV. **DISCUSSION**

30. In this section, we describe alternative methods of representing probabilistic structures and performing computation. In the first subsection, we describe how Bayesian Networks can be used to represent the data and allow filling-in data satisfying edit restraints. The advantage of Bayesian Networks is that there is considerable software that will automatically create probabilistic representations of the data with little or no modeling expertise. Because Bayesian Networks provide a very crude representation of the conditional probabilities needed for computing joint distributions, they will not perform as well as the method of Theorem 1. They can be expected to outperform many types of hot-deck imputation in terms of preserving joint distributions and satisfying edit restraints. In the second subsection, we describe extensions of Theorem 1 to situations where not all implicit edits can be generated.

IV.1 Much Simpler Procedures Using Bayesian Networks

31. Graphical representation of Bayes Nets and other probabilistic relationships date to Lauritzen and Spiegelhalter (1988). They are used extensively in machine learning. For instance, Getoor et al. (2001) demonstrate an efficient representation of Census data. 951 parameters are able to represent a potentially large number of cells in a contingency table (7 billion). Bayes Net software will quickly determine dependency relationship (see e.g. Figure 2 in Getoor et al. (2001). A mathematical representation is

$$P_B(A_1, \ldots, A_n) = \Pi_{i \leq n} \; P_B(A_i \mid \text{Parents}(A_i)). \qquad (1)$$

The advantage of the representation is that it allows extraordinarily fast computation of the probabilities of the form (1). There are many commercial and freeware software packages that automatically obtain a representation of the form (1), display the representation graphically, and give tools for easily modifying the representation to take advantage of different elementary dependency relationships. If a given variable depends on only a few other variables (i.e., the number of parents is small), then representation (1) is very efficient. If there is no missing data, then computation of the probabilities in (1) is exceedingly rapid (see e.g., Friedman 1997).

32 If we begin with either complete data or a combination of complete and incomplete data where fields are missing due to nonresponse or blanking due to edit restraints, then we can use

representation (1) to generate imputes for all of the missing data. The crucial advantage is that the representation is very parsimonious. It is very easy to generate imputations from the representation (1). A minor disadvantage is that we are only approximately generating the true underlying distributions within a factor of epsilon where epsilon may be much larger than the epsilon obtained when generating imputations using Theorem 1. More details of imputation procedures that use Bayes Networks are given in Thibaudeau and Winkler (2002).

IV.2 Alternate Computational Procedure for the Iterative Fitting and Imputation

33. In some situations such as very large labor force surveys with skip patterns, it is still not possible to generate all implicit edits. In those situations, Winkler and Chen (2002) provide a method for computing additional information about missing implicit edits "on-the-fly" when a large subset of all of the implicit edits is available. The method allows filling in most information (in the Fellegi-Holt OR sense) quickly using the available set of implicit edits. A few additional implicit associated with edit-failing records that cannot be error localized can be found via a algorithm that is much more efficient computationally than the cutting-plane algorithm 2 of Garfinkel, Kunnathur, and Liepins (1986). We note that we can fill in incomplete records using the method of Winkler and Chen (2002) but cannot put a suitable probabilistic structure on them.

34. We can adapt the information available from the representation $\boldsymbol{X}_C$ of Lemma 1 that does not use edit information. The EM fitting procedure gives a representation $\log(\boldsymbol{X}_C) = \sum_k a_k A_k$ (e.g., Bishop et al. 1975). The functions A_k are determined by the interactions that we are using. The coefficients a_k are obtained when we believe we have iterated sufficiently to get a solution that is close to the observed data according to our modeling criteria. We do not need to know the exact form of the a_k and A_k. The terms and coefficients are determined by an iterative fitting procedure such as multi-cycle ECM (MCECM) (Meng and Rubin 1993; also Winkler 1990, 1993). The additive representation has relationship to Hastie, Tibshirani, and Friedman (2001). A greedy function fitting method might be used as a more efficient computational alternative (Friedman 2001). As in classic loglinear modeling (e.g., Bishop et al. 1975), the terms a_k might be grouped into subgroups corresponding to different interactions. For instance, two-way interactions between fields y_1 and y_2 might correspond to probabilities $P(y_1 \in C, y_2 \in D)$, $P(y_1 \in C^c, y_2 \in D)$, $P(y_1 \in C, y_2 \in D^c)$ and $P(y_1 \in C^c, y_2 \in D^c)$.

35. Let r be a record in the last n-m rows of $\boldsymbol{X}$. Then, for convenience, we let $r = (x_{j1}, x_{j2}, \ldots, x_{jl}, \ldots.)$ where the last k-l columns must be filled-in. Using a Fellegi-Holt sequential fill-in procedure obtain N(j) new records $r_1, \ldots., r_{N(j)}$ corresponding to N(j) different ways that record r can be filled in. We need to sample from the model given by the probabilistic representation $\boldsymbol{X}_C$ of Lemma 1. That is, we sample a set of values (all at once) from one of the records in $\{\boldsymbol{X}_j^{c(i)}$ where $0 \le i \le N(j)\}$. To do this, we begin by assigning each record probability mass 1. Depending on the interactions present in each record, adjust the probability p_j associated with each record r_j according to the appropriate coefficients a_k. To select the actual filled-in record, sample record r_j, j = 1, …, N(j), with probability proportional to size p_j where the probabilities p_j are the appropriate cell probabilities from $\boldsymbol{X}_C$. If this is done for each of the last n-m rows of $\boldsymbol{X}$, then the resultant completed data $\boldsymbol{X}_D$ will have integer values, satisfy edit restraints, and preserve (approximately) the overall distribution determined by the model associated with $\boldsymbol{X}_C$.

36. We observe that the only rows of $\boldsymbol{X}_D$ where we know the appropriate probabilities of the cells are those that are filled in. The procedure adjusts the probabilities for the structural zeros in the rows where we fill in for missing data. It does this because each of the records that we match already satisfy edits.

V. CONCLUDING REMARKS

37. This paper provides an imputation methodology that generalizes hot-deck imputation. The methodology is consistent with the imputation method of Little and Rubin (1987) under the assumption of missing-at-random and with the edit-constraint model of Fellegi and Holt (1976). For large surveys with many edit restraints, the methods can be very computationally intensive. They target probability distributions and analytic concerns as exemplified in Little and Rubin.

1/ This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion. The authors thanks Dr. Yves Thibaudeau for comments on an earlier version of this paper.

REFERENCES

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

Chen, B.-C. (1998), "Set Covering Algorithms in Edit Generation," *American Statistical Association, Proceedings of the Section on Statistical Computing*, 91-96 (also available as Statistical Research Division Report rr98/06 at http://www.census.gov/srd/www/byyear.html).

Chen, B.-C., Thibaudeau, Y., and Winkler, W. (2002), "A Comparison Study of ACS If-Then-Else, NIM, and DISCRETE Edit and Imputation Systems Using ACS Data, *American Statistical Association, Proceedings of the Section on Survey Research Methods*, to appear (also longer research report at http://www.unece.org/stats/documents/2003/10/sde/wp.5.e.pdf).

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

Friedman, J. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, **29** (5), 1389-1432.

Friedman, N. (1997), "Learning Belief Networks in the Presence of Missing Values and Hidden Variables," *in D. Fisher, ed., `Proceedings of the Fourteenth International Conference on Machine Learning', Morgan Kaufmann, San Francisco, CA, pp. 125-133*.

Garfinkel, R. S., Kunnathur, A. S. and Liepins, G. E., (1986), "Optimal Imputation of Erroneous Data: Categorical Data, General Edits," *Operations Research*, **34**, 744-751.

Getoor, L., Taskar, B., and Koller, D. (2001), "Selectivity Estimation using Probabilistic Models," *Association of* Computing Machinery, Proceedings *of SIGMOD '01* (available at http://robotics.stanford.edu//~getoor/papers/sigmod01.ps).

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer: New York.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988), "Local Computations with Probabilities on Graphical Structures and Their application to Expert Systems," *JRSS, B 50(2)*, 157-224.

Little, R. A., and Rubin, D. B., (1987), *Statistical Analysis with Missing Data*, John Wiley: New York.

Meng, X.-L., and Rubin, D. B. (1993), "Maximum Likelihood Via the ECM Algorithm: A General Framework," *Biometrika*, **80**, 267-278.

Thibaudeau, Y., and Winkler, W. E. (2002), "Bayesian Network Representations, Generalized Imputation, and Synthetic Data Satisfying Analytic Restraints," (research report RRS2002/09 at http://www.census.gov/srd/www/byyear.html).

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 available as research report rr93/12 at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (1995), “Editing Discrete Data,” *American Statistical Association*, *Proceedings of the Section on Survey Research Methods*, 108-113 (also available as Statistical Research Division Report rr97/04 at http://www.census.gov/srd/www/byyear.html).

Winkler, W.E. (1997), “Set-Covering and Editing Discrete Data,” *American Statistical Association*, *Proceedings of the Section on Survey Research Methods*, 564-569 (also available as Statistical Research Division Report rr98/01at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. and Chen, B.-C. (2002), “Extending the Fellegi-Holt Model of Statistical Data Editing,” (research report RRS 2002/01 available at http://www.census.gov/srd/www/byyear.html).