Survey Methodology, June 1993
Vol. 19, No. 1, pp. 31–38
Statistics Canada

The Discrimination Power of Dependency Structures in Record Linkage

YVES THIBAUDEAU[1]

ABSTRACT

A record-linkage process brings together records from two files into pairs of two records, one from each file, for the purpose of comparison. Each record represents an individual. The status of the pair is a "matched pair" status if the two records in the pair represent the same individual. The status is an "unmatched pair" status if the two records do not represent the same individual. The record-linkage process is governed by an underlying probabilistic process. A record-linkage rule infers the status of each pair of records based on the value of the comparison. The pair is declared a "link" if the inferred status is that of a matched pair, and it is declared a "non-link" if the inferred status is that of an unmatched pair. The discrimination power of a record-linkage rule is the capacity of the rule to designate a maximum number of matched pairs as links, while keeping the rate of unmatched pairs designated as links to a minimum. In general, to construct a discriminatory record-linkage rule, some assumptions must be made on the structure of the underlying probabilistic process. In most of the existing literature, it is assumed that the underlying probabilistic process is an instance of the conditional independence latent class model. However, in many situations, this assumption is false. In fact, many underlying probabilistic processes do not exhibit key properties associated with conditional independence latent class models. The paper introduces more general models. In particular, latent class models with dependencies are studied and it is shown how they can improve the discrimination power of particular record-linkage rules.

KEY WORDS: Record-linkage rule; Latent class model; Expectation-Maximization procedures.

1. INTRODUCTION

The goal of the paper is to show how record-linkage rules can gain in discriminatory power when probabilistic models more descriptive of the underlying probabilistic processes, are elicited. For this purpose, a particular record-linkage situation is chosen and the conditional independence model, traditionally used in record linkage, is compared to a more descriptive model, in the sense that the new model allows for the expression of more complex relations of dependency between some of the variables involved.

First some terminology must be reviewed. In section 2, the definition of record-linkage process is stated and a general formulation of the probabilistic process underlying a record-linkage process is given. This formulation leads to the expression of two central concepts: the concepts of record-linkage rule and that of most discriminatory record-linkage rule.

In section 3, probabilistic models for record linkage are considered. In the first part of section 3, the family of latent class models is introduced and it is shown how this family provides natural models for the probabilistic process underlying a record-linkage process. In the second part of the section, the focus is on a particular model in the family of latent class models: the latent class model with conditional independence. This model is of interest because it is easy to handle computationally. In the third part, inference techniques adapted to the conditional independence model are reviewed.

In section 4, an application is presented. For this application, truth and falsehood are available, that is, it is known which pairs are matched and which aren't. The first part describes how the information on truth and falsehood was obtained. The second part shows how dependencies between the comparison fields are generated. In the third part of section 4, the knowledge on truth and falsehood is used to evaluate the dependencies between the comparison fields. This leads in the fourth part to the formulation of a model more descriptive of the underlying probabilistic structure of the record-linkage process. The final part is a brief discussion regarding the techniques of parameter estimation for generalized latent class models.

In section 5, an alternative methodology to construct approximate probabilistic models is presented. The model produced by this methodology is compared to those introduced in sections 3 and 4, in terms of discrimination power of the record-linkage rules derived from the models. The results of the comparisons are reported in section 6. In section 7, the suggestions of an anonymous referee to improve the methodology of the paper are presented. In section 8, conclusions are drawn and guidelines are provided.

[1] Yves Thibaudeau, U.S. Bureau of the Census, Federal Bldg. 4, Room 3000, Washington, D.C. 20233.

2. THE FELLEGI-SUNTER MODEL FOR RECORD-LINKAGE

2.1 Record-Linkage Processes

The paper is geared toward building new record-linkage techniques. Before expanding on new record-linkage techniques, some background is necessary. The concept of record-linkage process first needs to be reviewed. Consider two files; file A and file B, both containing records, each record representing an individual. A record-linkage process brings together one record from file A with one record from file B. The records are compared, producing the comparison pattern γ. For the purpose of this paper, this comparison pattern is a vector $\gamma = [\gamma^1, \ldots, \gamma^N]$, where N is the dimensionality of the vector. Each dimension corresponds to a comparison field recorded for each individual, such as last name, age, address, *etc.* With no loss of generality, γ^i is assigned the value 0 if the records disagree over comparison field i and it is assigned 1 if they agree. The comparison space Γ is assumed to be the set of all binary vectors (*i.e.* whose components are 0 or 1) of dimension N.

2.2 Underlying Probabilistic Processes

A record-linkage process is governed by an underlying probabilistic process. A good knowledge of the probabilistic process is needed to extract information from the record-linkage process. The formulation of the underlying probabilistic process is presented here in general terms. It is made more specific in the next section.

Consider a particular comparison pattern γ, define $m(\gamma)$ as the probability of observing γ, given that the two records producing γ, when brought together, represent the same individual. Similarly, define $u(\gamma)$ as the probability of observing γ, given that the two records producing γ, when brought together, do not represent the same individual. These two conditional probabilities, along with the probability of a match, define the underlying probabilistic process. The probabilistic process drives the record-linkage process. $m(\gamma)$ and $u(\gamma)$ are fundamental in the construction of record linkage rules; in particular most discriminatory rules. Record-linkage rules are devices to retrieve matches. They are defined next.

2.3 Record-Linkage Rules

In practice, a record-linkage rule classifies the pairs generated by a record-linkage process in one of three possible categories: a link, a non-link or a possible link. A link is an inferred matched pair and a non-link is an inferred unmatched pair. The pairs classified as possible links are set aside for further examination and eventually they are reclassified as links or non-links. The rule is based only on the value of the comparison vectors corresponding to each pair. The errors induced by a record-linkage rule are of two types: the type I error measuring the proportion of unmatched pairs among the pairs classified as links under the linkage rule, and the type II error measuring the proportion of matches among the pairs classified as non-links.

The objective of record-linkage, from the standpoint of the paper, is to construct a most discriminatory record-linkage rule; that is one that will retrieve a maximum number of links while keeping the type I error under control. To accomplish this, let the comparison patterns be indexed according to decreasing value of $m(\gamma)/u(\gamma)$ to obtain the sequence $\{\gamma_1, \gamma_2, \ldots \gamma_m\}$, where M is the total number of pairs. Fellegi and Sunter (1969) show that the rule declaring the pairs whose index is smaller than some upper bound K "links" is the most discriminatory record linkage rule. The upper bound K is a function of the maximum type I error tolerated. The rule is most discriminatory in the sense that for the same tolerance on the type I error, it is impossible to find another rule which, in the long run, will retrieve more matched pairs. This fact is a direct application of the Neyman-Pearson Lemma (DeGroot 1986, pp. 444-445). Two uses of the Fellegi-Sunter rule are illustrated in section 6.

The Fellegi-Sunter record-linkage rule is articulated around the ratio $m(\gamma)/u(\gamma)$. Usually this ratio is estimated from the data through a model of the underlying probabilistic process. It is assumed that the model is a genuine representation of the probabilistic process. If the representation is not genuine, then substituting $m(\gamma)/u(\gamma)$ in the Fellegi-Sunter rule may not yield a most discriminatory record-linkage rule. Therefore, particular care must be taken in the choice of the model. The next section introduces models designed to describe the underlying probabilistic process in given situations.

3. MODELS FOR RECORD-LINKAGE

Two models formulating underlying probabilistic processes are presented in this section. The first model is a general formulation of any underlying process. The second model is an application of the first. In some situations, the second model is a good representation of the underlying probabilistic process and the Fellegi-Sunter rule based on this model is most discriminatory. Parameter estimation is discussed so that the expressions involved in the Fellegi-Sunter rule can be evaluated.

3.1 Latent Class Models

Because of the particular nature of a record-linkage process, the underlying probabilistic process can always be represented by a latent class model. A latent class model is built around latent variables. Generally speaking, a latent variable is a variable not observable, characterizing any observation generated by the probabilistic process. Latent variables classify the observations into latent classes. In this problem, the observations are the comparison vectors (*i.e.* comparison patterns). An obvious latent variable categorizing the observations into two latent classes is the status of the pair associated with each comparison vector. This status is that of a matched pair status or of an unmatched pair status. The corresponding latent classes are the class of matched pairs and the class of unmatched pairs. A mathematical representation is given next to enable development of specific latent class models.

Let $\nu_{k,i_1,\ldots,i_N}$ represent the count of pairs with the following attributes: if $k = 0$ the corresponding pairs have an unmatched pair status and if $k = 1$ they have a matched pair status. Furthermore, whenever $i_s = 0$, the corresponding pairs do not exhibit record agreement over the comparison field s and whenever $i_s = 1$, the pairs do exhibit record agreement over the comparison field s. Note that $s = 1, \ldots, N$, where N is the number of comparison fields. It is important to keep in mind that the counts $\nu_{k,i_1}, \ldots, {}_{i_N}$ cannot be observed. Rather, what is observed are the counts aggregated over the latent classes. The aggregated counts are denoted by $\nu_{i_1}, \ldots, {}_{i_N}$, where

$$\nu_{i_1}, \ldots, {}_{i_N} = \nu_{0,i_1}, \ldots, {}_{i_N} + \nu_{1,i_1}, \ldots, {}_{i_N}. \tag{1}$$

While only the aggregated counts are observable in record-linkage situations, models are usually expressed in terms of the basic counts. This is done only for convenience. The following subsection is more specific and a simple latent class model for record linkage is introduced.

3.2 Conditional Independence

The conditional independence models are the simplest latent class models. Despite their simplicity, these models are an accurate representation of the underlying probabilistic process in some situations. Goodman (1974) gives a thorough analysis of several conditional independence models. Haberman (1979) gives a presentation of several conditional independence models, along with appropriate techniques of parameter estimation.

In this section, the conditional independence model for record linkage is introduced and its implications in terms of the underlying probabilistic process are exposed. The model is best described in its log-linear representation:

$$\log(\nu_{k,i_1}, \ldots, {}_{i_N}) = \mu + \lambda_k + \sum_{j=1}^{N} \alpha^j_{i_j} + \sum_{j=1}^{N} \zeta^j_{k,i_j}. \tag{2}$$

Naturally, there are constraints attached to the parameters of the model given in (2):

$$\lambda_1 = -\lambda_0;\ \alpha^j_1 = -\alpha^j_0;\ \zeta^j_{k,1} = -\zeta^j_{k,0};\ \zeta^j_{1,i_j} = -\zeta^j_{0,i_j}, \\ k = 0,1;\ j = 1, \ldots, N;\ i_j = 0,1. \tag{3}$$

The expression on the right-hand side of (2) includes one term for the latent variable (λ_k) and one term for each comparison field ($\alpha^j_{i_j}$). It also includes interaction terms (ζ^j_{k,i_j}). Each interaction is between a field and the latent variable. There are no direct interaction between the comparison fields. In other words, conditional on each latent class, agreements and disagreements over the comparison fields occur independently.

The assumption that the comparison variables are independent given the value of the latent variable is implicit when deriving inference through a conditional independence model. In practice, however, the underlying probabilistic process often conflicts with this assumption. Then the Fellegi-Sunter record-linkage rule constructed assuming model (2) may not be most discriminatory. In that situation, the discriminatory power can be raised through a better elicitation of the model. In fact, more elaborate latent class models integrate a higher degree of complexity in the relationships between the comparison fields themselves and between the comparison fields and the latent variable. These models can take a large number of forms according to the nature of a particular record-linkage situation. An instance of such models is presented in Section 4.

3.3 Parameter Estimation for the Conditional Independence Model

Once a model has been formulated, the values of its parameters must be evaluated. Then the Fellegi-Sunter rule is constructed from the model using the corresponding estimated values for $m(\gamma)$ and $u(\gamma)$. The parameter estimation process shall be reliable enough to prevent a significant loss of discriminatory power by way of the estimation error.

One feature of the latent class models makes them prone to estimation error: unidentifiability. Latent class models typically are unidentifiable in the sense that the equations maximizing the likelihood admit more than one solution. Parameter estimation with unidentifiable models remains difficult and confusing. However, from experience, the author found that for the conditional independence models, unidentifiability is usually not a determinant factor in the estimation error. A larger part of the error typically comes from the inadequacy of the model as a genuine representation of the underlying probabilistic process.

A suitable parameter-estimation technique for conditional independence models stems from approaching the problem as one of finding a maximum likelihood estimator in the presence of "missing observations". The missing observation in this case is the latent variable, the status of each pair. In the general context of parameter estimation with missing observations, Expectation-Maximization (E.M.) algorithms are quite popular. In fact, the E.M. algorithm is implemented without difficulty in the estimation of the parameters of the conditional independence model given in (2) (Winkler 1988). But if there is considerable departure from the independence assumption, the value of the estimates becomes difficult to interpret (An example of this is given in section 4).

4. THE ST. LOUIS DATA: AN EXAMPLE OF A COMPLEX RECORD-LINKAGE PROCESS

This section introduces a particular example of a record-linkage process. A model is developed specifically to represent the underlying probabilistic process supporting this record-linkage process. It is expected that this model will induce more discrimination power in the application of the Fellegi-Sunter rule than the conditional independence model would.

4.1 Observable Latent Variable

The example is based on data collected in 1988 during a dress rehearsal in preparation for the Decennial Census Operations. Basically, there are two separate and presumably exhaustive surveys of all the individuals living in a defined geographical area within the city of St. Louis, Missouri. For each survey and for each individual available at the time of the survey, a record is created and various characteristics of the individual are recorded. These characteristics are: house number, phone number, street name, first name, last name, middle initial, marital status, age, race, sex, relationship with the respondent. The records of the two surveys are linked together.

For this particular application, the latent variable is made observable through an extensive follow-up study for the purpose of this and other researches. In the present situation, the information extracted from the latent variable leads to the construction of a model representative of the probabilistic process underlying the record-linkage process. Ultimately the discrimination power of this model is compared with that of the conditional independence model. The motivations leading to the construction of the model are presented in the following subsections.

4.2 Blocking and Dependencies

The goal of record-linkage is to retrieve as many matched pairs as possible given an upper bound on the type I error. The first obstacle is often the size of the files. The files may be quite large, making it impossible to examine all the pairs consisting of one record of file A and one record of file B. Blocking is considered whenever an exhaustive review of all the pairs is too costly and/or too time consuming.

The principle of blocking is as follows: To bring down the number of comparisons and other associated operations, the records of each file are assigned to blocks according to the value of a few key characteristics. These characteristics are called the blocking variables. Only the records whose blocking variables take the same values may be brought into pairs. Since the records forming a matched pair tend to agree on the blocking characteristics, it is natural to expect the vast majority of the pairs discarded to be unmatched, as a result of the blocking scheme.

In the St. Louis example, the census file has 15,048 records, while the PES file contains 12,072 records. Potentially, there are over 180,000,000 pairs available for review. This number is excessive and blocking must be used to keep the size of the problem manageable. Therefore, the records are blocked on the first character of the surname and on a geographical unit called geocode. The geographical area encompassed by a given value of the geocode may consist of several street blocks, or two or more nearby perpendicular or parallel streets. This scheme yields blocks of reasonable sizes. Under this design, 116,305 pairs provide the information to construct inference.

Unfortunately, while it brings down the size of the problem, blocking on geocode also has undesirable side effects: it induces strong dependencies between the household variables among the unmatched pairs. The household variables are the last name, house number, street name and telephone number. For instance, consider two individuals forming an unmatched pair but who are part of the same block. Now, suppose these two individuals agree on the last name. Intuitively, given this information, chances are higher that the two individuals are from the same household. Therefore, the probabilities of agreement over the other household fields, given the information of agreement on the last name, are higher than the marginal probabilities. The nature of the dependencies between the household variables is studied next.

4.3 Measuring the Dependencies

To construct a model representative of the St. Louis record-linkage process, the dependencies between the household variables must be assessed. The information on the latent variable allows this. Table 1 gives the correlations of the responses of record comparisons over the comparison fields for the matched pairs. Table 2 gives the correlations of the responses of the record comparisons over the comparison fields for the unmatched pairs. For both matrices, all the correlations greater or equal to .01 are given. A correlation is not shown only if it is smaller than .01.

The correlations in Table 1, are rather small and overall do not suggest a significant pattern of dependency among the comparison variables restricted to the matched pairs. Note in particular that the correlations between the household variables are small among the matched pairs, suggesting little or no dependency. This can be explained by the fact that among the matched pairs, the agreement rate over any household field is very high and has a behavior close to that of a constant.

Table 1

Correlations Between Selected Comparison Fields over the Set of Links

	Middle In.	Street	Phone	Marital
First Name	.123	0.	.045	.032
Middle In.	1	.010	.161	.079
House No.	.017	.194	.037	0.
Street	.01	1	.035	0.
Phone	.161	.035	1	.107
Age	.051	.004	.075	.118
Marital	.079	0.	.107	1

Table 2

Correlations Between Selected Comparison Fields over the Set of Non-Links

	House No.	Street	Phone	Marital	Race
Last N.	.748	.326	.642	.099	.101
House No.	1	.400	.699	.111	.105
Street	.400	1	.292	.043	.086
Age	.104	.054	.086	.165	.024
Rel	.121	.068	.084	.394	.049

But in Table 2, the effects of blocking are evident in the high values of the correlations associated with the household variables restricted to the unmatched pairs. A sensible design for the model of the underlying probabilistic process should account for these high correlations by incorporating dependency components.

4.4 A Model Tailored for the St. Louis Data

In order to make valid inference on the status of the pairs, a model descriptive of the underlying probabilistic process must be elicited. The conditional independence model presented in (2) is attractive because of its simplicity. However, it is clear at this point that this model does not correctly represent the probabilistic process underlying the St. Louis record-linkage process. An educated model is introduced, motivated by the information made available on the dependencies between the household variables.

To appreciate the more general structure of the educated model, some conventions must be set regarding the indexing of the comparison fields: comparison field 1 is the last name, comparison field 2 is the house number, comparison field 3 is the street name, and comparison field 4 is the phone number. The seven remaining comparison fields are indexed arbitrarily by the values 5-11. The educated model accounts for all possible interaction effects between fields 1 through 4 among the unmatched pairs. The log-linear representation of the educated model is as follows:

$$\log(\nu_{k,i_{1,\ldots,}i_{11}}) = \mu + \lambda_k + \sum_{j=1}^{11} \alpha^{j}_{i_j} + \sum_{j=1}^{11} \zeta^{j}_{k,i_j}$$

$$+ (1-k)\left(\sum_{\{\le j<l\le 4\}} \eta^{j,l}_{i_j,i_l} \right.$$

$$\left. + \sum_{\{1\le j<l<m\le 4\}} \Phi^{j,l,m}_{i_j,i_l,i_m} + \Psi^{1,2,3,4}_{i_1,i_2,i_3,i_4} \right). \qquad (4)$$

Note the coefficient $(1-k)$ multiplying the household interaction terms, indicating that the dependency relation between the household variables is only among the unmatched pairs. This contrasts with the symmetry of the conditional independence model in (2).

The restrictions in (3) apply here as well. In addition, more constraints must be satisfied. The following constraints are imposed on the interaction terms of the second order:

$$\eta^{j,l}_{i_j,1} = -\eta^{j,l}_{i_j,0}; \quad \eta^{j,l}_{1,i_l} = -\eta^{j,l}_{0,i_l}. \qquad (5)$$

The range of the indices is $1 \le j < l \le 4$. The constraints on the interaction terms of the third order are:

$$\Phi^{j,l,m}_{i_j,i_l,1} = -\Phi^{j,l,m}_{i_j,i_l,0}; \quad \Phi^{j,l,m}_{i_j,1,i_m} = -\Phi^{j,l,m}_{i_j,0,i_m};$$

$$\Phi^{j,l,m}_{1,i_l,i_m} = -\Phi^{j,l,m}_{0,i_l,i_m}. \qquad (6)$$

The range of the indices in this case is: $1 \le j < l < m \le 4$. Finally, the constraints on the fourth order interaction terms are:

$$\Psi^{1,2,3,4}_{i_1,i_2,i_3,1} = -\Psi^{1,2,3,4}_{i_1,i_2,i_3,0}; \quad \Psi^{1,2,3,4}_{i_1,i_2,1,i_4} = -\Psi^{1,2,3,4}_{i_1,i_2,0,i_4};$$

$$\Psi^{1,2,3,4}_{i_1,1,i_3,i_4} = -\Psi^{1,2,3,4}_{i_1,0,i_3,i_4}; \quad \Psi^{1,2,3,4}_{1,i_2,i_3,i_4} = -\Psi^{1,2,3,4}_{0,i_2,i_3,i_4}. \qquad (7)$$

It is natural to expect the educated model (4) to be more discriminatory since it accounts for interactions between the household variables. In section 6, the performances of the two models are presented.

4.5 Parameter Estimation for Models with Dependencies

Parameter estimation for models with dependencies is far more difficult than for conditional independence models. For the St. Louis example, the scoring algorithm given by Haberman (1979, p. 547) was used to estimate the parameters of the educated model (4). This technique can be regarded as an E.M. algorithm where the maximization part (M. step) is an application of the Newton-Ralphson algorithm.

The most important difficulty when using this technique is the choice of a starting point. The following strategy is adopted to choose a starting point. First, the parameters of the conditional independence model (2) are estimated via the E.M. algorithm presented in subsection 3.3. Then an intermediate model is constructed. The intermediate model, in this case, embeds all the second and lower order interaction terms of the educated model (4). The estimated parameters of the conditional independence model can serve to construct the starting point to estimate the parameters of the intermediate model through the scoring algorithm. Finally, the estimates of the parameters of the intermediate model are used as a starting point to estimate the parameters of the educated model (4), via the scoring algorithm.

5. THE AD-HOC APPROACH

In the last section, a complex model representing an underlying probabilistic process was elicited for the St. Louis data. In this situation, the elicitation is easy since follow-up information is available. Of course in practice, follow-up information is not available. It is often too difficult and/or too expensive to go through the elicitation and estimation procedures to determine the structure of the underlying process and the values of the parameters. In those cases, an ad-hoc approach might be appropriate. In the St. Louis example, the ad-hoc approach consists of adjusting the parameters of the process derived from the conditional independence model (2) to obtain a more discriminatory model.

Note that under both model (2) and model (4), for the matched pair, the agreement or disagreements over the comparison fields are independent. This means that the following formula applies in both situations.

$$m(\gamma) = \prod_{i-1}^{N} m_i^{x_i}(1 - m_i)^{1-x_i},$$

m_i is the probability of agreement over field i of two records forming a matched pair. Furthermore, $x_i = 0$ if the pattern γ calls for a disagreement over field i and $x_i = 1$ if it calls for an agreement. The idea behind the ad-hoc method is to keep the conditional independence structure in (2), but to adjust the values of the m_i's.

The probabilities of agreement, conditional on a matched pair, evaluated under the conditional independence model and the educated model are given in Table 3. The difference between the probability corresponding to the educated model with the probability corresponding to the conditional independence model can be quite substantial for some fields. In particular, the difference is important in the case of the first name field.

Table 3

Probabilities of Agreement Conditional on a Matched Pair

Comparison Field	Cond'l Indep.	Educated
Last Name	.9430	.9561
First Name	.3319	.9140
Mid. Init.	.2125	.5222
House No.	.9692	.9724
Street Name	.9179	.9194
Phone	.6619	.6887
Age	.3903	.8602
Relation	.3353	.4986
Marital Status	.6072	.8547
Sex	.6134	.4842
Race	.9672	.9018

In general, experience shows that the conditional probability of agreement over first name, conditional on a matched pair, is around .99, closer to the .91 value obtained under the educated model. Therefore, after estimating the parameters of the conditional independence model through the E.M. algorithm, the probability of agreement over the first name given a match status is replaced by the value .99. The probability of agreement over the last name given a matched pair is also replaced by the value .99. This procedure increases the discriminatory power associated with the conditional independence model in the application of the Fellegi-Sunter rule.

6. APPLYING THE FELLEGI-SUNTER RULE

6.1 St.Louis

This subsection evaluates the discrimination power of the Fellegi-Sunter rule when applied to the St-Louis record-linkage data and assuming, in turn, three different underlying probabilistic processes. The three underlying probabilistic processes assumed are derived directly from the conditional model (2), directly from the educated model (4), and finally, from the conditional model (2), through the ad-hoc procedure. The following table gives a comparative measure of the performance of the Fellegi-Sunter rule under each of the 3 assumptions regarding the underlying process. The performance is evaluated making use of the privileged information available on the latent variable.

Each cell of Table 4 contains three entries. The first of these entries is the number of matched pairs that were designated links through the Fellegi-Sunter record-linkage rule, assuming each of the three underlying processes, and under four different controlled Type I errors. The total number of matched pairs that could theoretically be recovered is 9,823. The second entry of each cell is the total number of pairs designated link through the Fellegi-Sunter rule. The third entry of the cell is the upper bound on the

Type I error. Recall that the Fellegi-Sunter rule maximizes the number of links under a fixed type I error provided it is based on the correct underlying process. The first column of Table 4 gives the counts assuming an underlying process derived from the conditional independence model (2). The second column gives the same quantities assuming an underlying process derived from the educated model (4). Finally, the third column gives the same numbers assuming an underlying process derived from the conditional independence model and adjusted through the ad-hoc procedure.

Table 4

St. Louis: Links Recovered via Three Approaches under Four Error Levels

	Independence Assumption	Household Interactions	Ad-hoc Procedure
Links	6,404	9,012	6,476
Pairs	6,436	9,056	6,508
Error Bound	.005	.005	.005
Links	7,273	9,712	9,562
Pairs	7,346	9,808	9,659
Error Bound	.01	.01	.01
Links	9,636	9,758	9,765
Pairs	9,824	9,952	9,960
Error Bound	.02	.02	.02
Links	9,740	9,776	9,783
Pairs	10,038	10,062	10,097
Error Bound	.03	.03	.03

There are two important facts that can be deduced from this table. First, the rule based on an underlying process derived from the educated model (4) does consistently better than the rule based on an underlying process derived from the conditional independence model in terms of matches retrieved. Secondly, the performances of the rules differ most when the bound on the type I error is small and at that level (.005), the rule based on an underlying probability process derived from the educated model is clearly superior. When the bound is larger (.03), the underlying probabilistic models are more or less equivalent in terms of induced discrimination power.

6.2 Columbia

The same type of data were collected throughout the area of Columbia, Missouri. The data are slightly different because some of the records have a rural format, that is the street name is replaced by the rural route number and the house number by the box number. Nevertheless, the same relations of dependencies emerge and the same model is appropriate. Table 5 gives a summary of the discrimination achieved at 2 levels of tolerance on the type I error. Taking into account the blocking scheme, there are 6,780 retrievable pairs.

Table 5

Columbia: Links Recovered via Three Approaches under Two Error Levels

	Independence Assumption	Household Interactions	Ad-hoc Procedure
Links	700	1,268	2,035
Pairs	704	1,276	2,046
Type I Error	.005	.005	.005
Links	5,954	6,607	6,545
Pairs	6,016	6,675	6,612
Type I Error	.01	.01	.01

In the case of Columbia, it is clear again than the educated model does better than the conditional independence model. It should be noted that in practice, the ad-hoc approach built on the conditional independence model performs as well as the educated model. The educated model however, is preferred because of its sound theoretical basis.

7. A SUGGESTION FROM AN ANONYMOUS REFEREEE

Another ad-hoc technique is suggested by an anonymous referee. The referee points out that a large majority of the pairs examined in situations like these are unmatched. In the case of St. Louis, 91.5% of the pairs examined turn out to be unmatched. Given this proportion, the trends animating the comparison variables over the set of all pairs, mostly reflect the activity of the unmatched pairs. This reasoning can be extended further to conclude that the estimation of the parameters of the dependency structure underlying the unmatched pairs can be carried through successfully by treating the set of all pairs as if it were the set of unmatched pairs. The parameter estimation becomes trivial. The parameters that must be estimated characterize a simple log-linear model, without any latent variable (Fienberg, Bishop and Holland, p. 24). The parameters descriptive of the matches can be estimated separately through a simple iterative technique such as the E.M. algorithm, combined with *a priori* information.

The approach of the referee does proceed from a realistic model of the process, and in that way, it is in agreement with the thrust of this paper. But the effort of the paper is also to devise discriminatory rules, while sticking to the latent structure constraint. In situations where the proportion of matched pairs is high, or dependencies are manifest among the matches, the approach of the referee fails. A parameter estimation derived directly from the natural model, if feasible, is recommended.

8. CONCLUSIONS

The goal of the research was to show how a better elicitation of the probabilistic models supporting record-linkage processes can induce accrued discriminatory power in the Fellegi-Sunter record-linkage rule. In the cases of the St. Louis and Columbia examples, this goal was certainly achieved. The educated model given in (4) is indeed more descriptive of the underlying probabilistic process and it induces a good deal more discrimination power in the Fellegi-Sunter rule than the conditional independence model (2).

The techniques used for the St. Louis and Columbia data can also be used for the analysis of other data set generated by record-linkage processes supported by a probabilistic process with a similar dependency structure. This dependency structure is certain to surface in any record-linkage application involving the matching of records of individuals on a set of household variables (last name, street name, house number, phone, rural address *etc.*). It is also likely to occur when matching records of businesses on household variables.

There are two major difficulties in the way, when seeking improved discriminatory power by model elicitation. First, since the probability structure underlying the process is usually unknown, to elicit the structure or the corresponding statistical model involves a considerable investigative effort and the cost involved may be prohibitive. Second, even assuming that the correct model is available, the estimation procedures available for the parameter estimation are difficult to handle and poorly understood. More research and work are needed to understand and, to a degree, overcome these two difficulties.

It must also be pointed out that methods based on ad-hoc adjustments of the type described in section 5, and on approximations, as suggested by an anonymous referee, also increase the discriminatory power of the Fellegi-Sunter rule substantially in situations of the type of St. Louis or Columbia. Techniques of this type are serious competitors. The parameter estimation is easy and the associated Fellegi-Sunter rule can be just as crisp in some cases. However, the assumptions supporting these techniques are flawed and the resulting Fellegi-Sunter rule is pathological, providing an unsteady basis on which to make decisions. A model with parameters estimated ''naturally'' is preferable. The ad-hoc techniques and approximations are recommended when the elicitation of an educated model seems not possible, or the estimation of the parameters of the educated model appears excessively difficult.

A word must be said about the St. Louis and Columbia data. These data are of very high quality. This explains in part the very successful rate of matching exhibited in both the St. Louis and Columbia examples. It is also reasonable to expect a less clear-cut difference between the various linkage techniques had the data been lower quality.

ACKNOWLEDGEMENTS

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau. The author is grateful to the anonymous referee for his/her patience and constructive suggestions. The author is also indebted to William E. Winkler for the guidance he provided throughout the learning process that led to this paper.

REFERENCES

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: MIT Press.

DeGROOT, M.H. (1986). *Probability and Statistics*, 2nd. Edition. Reading, MA: Addison-Wesley.

FELLEGI, I.P., and SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 40, 1183-1210.

GOODMAN, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 2, 215-231.

HABERMAN, S.J. (1979). *Analysis of Qualitative Data*, Vol. 2. New York: Academic Press.

HABERMAN, S.J. (1976). Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation. *Proceedings of the Statistical Computing Section, American Statistical Association*, 45-50.

THIBAUDEAU, Y. (1989). Fitting log-linear models in computer matching. *Proceedings of the Statistical Computing Section, American Statistical Association*, 283-288.

WINKLER, W.E. (1989). Near automatic weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Bureau of The Census Fifth Annual Research Conference*, 145-155.

WINKLER, W.E. (1988). Using The E.M. algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 667-671.