Posterior Prototyping: Bridging the Gap between Bayesian Record Linkage and Regression

Andee Kaplan*¹, Brenda Betancourt² and Rebecca C. Steorts¹

¹Department of Statistical Science, Duke University ²Department of Statistics, University of Florida

October 12, 2018

Abstract

Record linkage (entity resolution or de-deduplication) is the process of merging noisy databases to remove duplicate entities. While record linkage removes duplicate entities from the data, many researchers are interested in performing inference, prediction or post-linkage analysis on the linked data, which we call the *downstream task*. Depending on the downstream task, one may wish to find the most representative record before performing the post-linkage analysis. Motivated by the downstream task, we propose first performing record linkage using a Bayesian model and then choosing representative records through *prototyping*. Given the information about the representative records, we then explore two downstream tasks — linear regression and binary classification via logistic regression. In addition, we explore how error propagation occurs in both of these settings. We provide thorough empirical studies for our proposed methodology, and conclude with a discussion of practical insights into our work.

1 Introduction

Record linkage (entity resolution or de-duplication) is used to join multiple databases to remove duplicate entities. While record linkage removes the duplicate entities from the data, many researchers are interested in performing inference, prediction, or post-linkage analysis on the linked data (e.g., regression or capture-recapture), which we call the downstream task. Depending on the downstream task, one may wish to find the most representative record before performing the post-linkage analysis. For example, when the values of features used in a downstream task differ for linked data, which values should be used? To motivate this more clearly, consider modeling blood pressure (bp) using the following two features (covariates): income and sex. In addition, we assume that we perform this task after performing record linkage using the following features: first and last name and full data of birth. To further illustrate this motivational scenario, we provide an example of four records that are thought to represent the same individual after performing a record linkage process (see Table 1). Examination of this table raises important questions that need to be addressed before performing a particular downstream task, such as which values of bp, income, and sex should be used as the representative features (or covariates) in a regression model? The goal of this paper is to provide viable solutions to this question, with a guidance for the choice of the best approach based on downstream performance and error propagation.

Methods for the analysis of linked data have been numerous in recent years. However, most approaches have been limited primarily to two-file matching. For example, [14] addressed the problem of linking two databases under the assumption that they represent a permutation of the same set of records and the linkage error only involves the response variable. They proposed an unbiased estimator (LL) for linear regression, conditional on the matching probabilities provided by the linkage process. [10] extended [14] to handle more realistic linkage scenarios under a logistic regression framework. Generalizations of the LL estimator

^{*}We thank Neil Marchant, Ben Rubenstein, and the Australian Bureau of Statistics for providing the package eber.

First	Last	Birthdate	Sex	Education	Income	BP	high_bp
nixolas	re8d	1985-8-19	M	Advanced degree	79	158	1
nicholast	relr	1985 - 8 - 22	\mathbf{M}	Some college or associate degree	79	149	1
riicholaz	rid	1985 - 8 - 17	\mathbf{M}	Less than a high school diploma	44	131	1
nicholas	reid	1985 - 8 - 22	\mathbf{M}	Advanced degree	79	131	1

Table 1: Records that represent the same entity according to a record linkage task.

can be found in [13], where estimating equations provide consistent estimators of population quantities. [8] relaxed these assumptions and considered the matching probabilities as prior information to be used within a multiple imputation scenario. The previously mentioned approaches follow a two-stage modeling framework where the matching probabilities of record pairs provided by the record linkage model, are later introduced in the regression modeling strategy. Alternatively, single-stage approaches for the two-file case that jointly model the record linkage task and the association between key variables have been proposed by [11] for survival data using a frequentist procedure, while [9] and [5] proposed Bayesian methods for regression in a medical application and a general setting, respectively.

In contrast to the existing literature, we are interested in the general case of linking multiple databases assuming that both the response and predictor variables are susceptible to linkage error in the context of regression. The proposed procedure can be thought of as a middle step to facilitate the transition from linked data to post-linkage analysis in a two-stage modeling framework. A significant advantage of our proposal compared to existing approaches is that inference for regression (and other post-linkage analyses) can be performed in a traditional manner after the representative data set is constructed through prototyping. In this way, the information from the record linkage process is transferred through to the analysis stage allowing for uncertainty propagation. Bayesian methods have a long history of use in record linkage due to their flexibility and exact error propagation but only recently some procedures have been proposed to deal with more than two files [18, 22, 24].

The paper is organized as follows. Section 2 introduces the notation and empirical Bayesian graphical record linkage model of [24, 21, 22] that is used throughout the rest of the paper. Section 3 details two downstream tasks that are commonly assumed in the literature [14, 13, 8], where we contrast previously explored settings with the assumptions in our paper. Here, we propose four methods for selecting a representative set of records for use in the downstream tasks, which are intuitive and simple. Section 4 provides two empirical studies of our proposed methodology regarding the downstream task. Section 5 provides some guidelines for practitioners looking to perform prototyping prior to completing downstream analyses and also provides a discussion of our results and directions for future work.

2 Bayesian Record Linkage

While our framework for prototyping is not tied to any one particular record linkage method, we use the framework of [22], which clusters similar records to a latent entity that represents the true record. This approach is based on empirical Bayesian principles and allows both categorical and string-valued variables. The flexibility of the model and existence of provable performance bounds ([23]) make this modeling approach suitable for practical applications. Finally, this method is easy to implement as software is publically available on CRAN.

In the approach of [22], noisy data are assumed organized into multiple databases (or lists) and each record consists of multiple fields (or attributes). We consider each record as a representation of some true individual, who is not observed, but rather is considered latent. Let $X_{ij\ell}$ denote the observed value of the ℓ th field for the jth record in the ith list, and $Y_{j'\ell}$ denotes the true value of the ℓ th field for the jth latent entity. Then Λ_{ij} denotes the latent entity to which the jth record in the ith list corresponds, i.e., $X_{ij\ell}$ and $Y_{j'\ell}$ represent the same entity if and only if $\Lambda_{ij} = j'$. Denote distortion by $z_{ij\ell} = I(X_{ij\ell} \neq Y_{\Lambda_{ij\ell}})$, where $I(\cdot)$ represents the indicator function. Let δ_a denote the distribution of a point mass at a (e.g., $\delta_{y_{\Lambda_{ij\ell}}}$). [22] assumed fields $1, \ldots, p_s$ are string-valued, while fields $p_s + 1, \ldots, p_s + p_c$ are categorical, where $p_s + p_c = p$ is the total number of fields. Additionally, they assumed an empirical Bayesian distribution on the latent parameter. For each $\ell \in \{1, \ldots, p_s + p_c\}$, let S_ℓ represent the set of all values for the ℓ th field that occurs in

the data, i.e., $S_{\ell} = \{X_{ij\ell} : 1 \leq i \leq k, 1 \leq j \leq n_i\}$, and let $\alpha_{\ell}(w)$ equal the empirical frequency of value w in field ℓ . Let G_{ℓ} denote the empirical distribution of the data in the ℓ th field from all records in all databases. So, if a random variable W has distribution G_{ℓ} , then for every $w \in S_{\ell}$, $P(W = w) = \alpha_{\ell}(w)$. Hence, let G_{ℓ} be the prior for each latent entity $Y_{j'\ell}$. The distortion process is then defined to be

$$F_{\ell}(Y_{\Lambda_{ij}\ell}) = P(X_{ij\ell} = w \mid \Lambda_{ij}, Y_{\Lambda_{ij}\ell}, z_{ij\ell}) = \frac{\alpha_{\ell}(w) \exp[-c d(w, Y_{\Lambda_{ij}\ell})]}{\sum_{w \in S_{\ell}} \alpha_{\ell}(w) \exp[-c d(w, Y_{\Lambda_{ij}\ell})]},$$

where c > 0 is a fixed normalizing constant corresponding to an arbitrary distance metric $d(\cdot, \cdot)$. The full record linkage model specification can be written as follows.

$$X_{ij\ell} \mid \Lambda_{ij}, Y_{\Lambda_{ij}\ell}, z_{ij\ell} \stackrel{\text{ind}}{\sim} \begin{cases} \delta(Y_{\Lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 0 \\ F_{\ell}(Y_{\Lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 1, \ell \leq p_s \\ G_{\ell} & \text{if } z_{ij\ell} = 1, \ell > p_s \end{cases}$$

$$Y_{j'\ell} \stackrel{\text{ind}}{\sim} G_{\ell}$$

$$z_{ij\ell} \mid \beta_{i\ell} \stackrel{\text{ind}}{\sim} \text{Bernoulli}(\beta_{i\ell}), \qquad (2.1)$$

$$\beta_{i\ell} \stackrel{\text{ind}}{\sim} \text{Beta}(a, b)$$

$$\Lambda_{ij} \stackrel{\text{ind}}{\sim} \text{Uniform } (1, \dots, M), \qquad (2.2)$$

where all distributions are independent of each other, and the parameters a, b, M are assumed known. The collection of the cluster assignments Λ_{ij} is denoted by Λ and represents the linkage structure of the data.

3 Representative Records and Downstream Tasks

Principled approaches that perform linkage and downstream tasks jointly using a fully Bayesian framework have the advantage of error propagation but are challenging to implement in big data scenarios, and the issue of differing attribute values for linked records remains. We introduce *prototyping* as an alternative to ameliorate computational costs while preserving accuracy in the downstream task. Given the many ways that one could choose a representative record, we propose four methods for this task (Section 3.1). We discuss their potential benefits and risks before examining them in an empirical setting (Section 4). We then discuss two downstream tasks (Section 3.2) — linear regression and binary classification — that are used in our empirical studies.

3.1 Representative Records

To bridge the gap between record linkage and the downstream task, we propose four methods to choose or create the representative record from linked data. This process is a function of the data and the linkage structure, and we present both probabilistic and deterministic functions. The result in all cases is a representative data set to be passed on to the downstream task.

3.1.1 Random Prototyping

Our first proposal to choose a representative record (prototype) for a cluster is the simplest and serves as a baseline or benchmark. One simply chooses the representative record uniformly at random or using a more informed distribution. More specifically, we propose either choosing the record record uniformly at random or using the pairwise posterior linkage probabilities resulting from model (2.2) to inform the choice.

3.1.2 Minimax Prototyping

Our second proposal to choose a representative record is to select the record that "most closely captures" that of the latent entity. Of course, this is quite subjective. We propose selecting the record whose farthest

neighbors within the cluster is closest, where closeness is measured by a record distance function, $d_r(\cdot)$. We can write this as the record r = (i, j) within each cluster $\Lambda_{j'}$ such that

$$r = \arg\min_{(i,j) \in \Lambda_{j'}} \max_{(i^*,j^*) \in \Lambda_{j'}} d_r((i,j),(i^*,j^*)).$$

The result is a set of representative records, one for each latent individual, that is closest to the other records in each cluster. When there is a tie within the cluster, we select a record uniformly at random.

There are many distance functions that can be used for $d_r(\cdot,\cdot)$. We define the distance function to be a weighted average of individual variable-level distances that depend on the column type. Given two records, (i,j) and (i*,j*), we use a weighted average of column-wise distances (based on the column type) to produce the following single distance metric:

$$d_r((i,j),(i*,j*)) = \sum_{\ell=1}^p w_\ell d_{r\ell}((i,j),(i^*,j^*)),$$

where $\sum_{\ell=1}^{p} w_{\ell} = 1$. The column-wise distance functions $d_{r\ell}(\cdot, \cdot)$ we use are presented in Table 2. The weighting

Column	$d_{r\ell}(\cdot,\cdot)$
String	Any string distance function, i.e. Jaro-Winkler string distance [31].
Numeric	Absolute distance, $d_{r\ell}((i,j),(i^*,j^*)) = x_{ij\ell} - x_{i^*j^*\ell} $
Categorical	Binary distance, $d_{r\ell}((i,j),(i^*,j^*)) = \mathbb{I}(x_{ij\ell}! = x_{i^*j^*\ell})$
Ordinal	Absolute distance between levels. Let $\gamma(x_{ij\ell})$ be the order of the
	value $x_{ij\ell}$, then $d_{r\ell}((i,j),(i^*,j^*)) = \gamma(x_{ij\ell}) - \gamma(x_{i^*j^*\ell}) $

Table 2: Column-wise distance functions based on column type used to create a distance metric between two records.

of variable distances is used to place importance on individual features according to prior knowledge of the data set and to scale the feature distances to a common range. In this paper, we scale all column-wise distances to be values between 0 and 1.

3.1.3 Composite Records

Our third proposal to choose a representative record is by aggregating the records (in each cluster) to form a composite record that includes information from each linked record. The form of aggregation can depend on the column type, and the aggregation itself can be weighted by some prior knowledge of the data sources or use the posterior information from model (2.2). For quantitative variables, we use a weighted arithmetic mean to combine linked values, whereas for categorical variables, a weighted majority vote is used. For string variables, we use a weighted majority vote for each character, which allows for noisy strings to differ on a continuum.

3.1.4 Posterior Prototyping

Our fourth proposal to choose a representative record utilizes the minimax prototyping method in a fully Bayesian setting. This is desirable as the posterior distribution of the linkage is used to weight the downstream tasks, which allows the error from the record linkage task to be naturally propagated into the downstream task.

We propose two methods for utilizing the posterior prototyping (PP) weights — a weighted downstream task and a thresholded representative data set based on the weights. As already mentioned, PP weights naturally propagate the linkage error into the downstream task, which we now explain. For each MCMC iteration from the Bayesian record linkage model, we obtain the most representative records using minimax prototyping and then compute the probability of each record being selected over all MCMC iterations. The

¹This is closely related to the use of minimax linkages used in hierarchical clustering tasks [3].

posterior prototyping (PP) probabilities can then either be used as weights for each record in the regression or as a thresholded variant where we only include records whose PP weights are above 0.5^2 .

These four proposed methods each have potential benefits. The goal of prototyping is to select the correct representations of latent entities as often as possible; however, uniform random selection has no means to achieve this goal. Turning to minimax selection, if a distance function can accurately reflect the distance between pairs of records in the data set, then this method may perform well (see Sections 4.2-4.3.3 for evidence). Alternatively, composite records necessarily alter the data for all entities with multiple copies in the data, affecting some downstream tasks (like linear regression) heavily. The ability of posterior prototyping to propagate record linkage error to the downstream task is an attractive feature and a great strength of the Bayesian paradigm. In addition, the ability to use the entire posterior distribution of the linkage structure also poses the potential for superior downstream performance (See Section 4.3.2 for evidence).

3.2 Downstream Tasks

A downstream task can be any model or exploratory analysis performed on the data set after the linkage step. We consider two commonly used downstream tasks — linear regression and binary classification via logistic regression. There are two cases that can be considered for the downstream task with respect to error propagation. First, the features involved in the downstream task are only present in one database. In this scenario the downstream task is straightforward as the error can be propagated exactly (see [14, 13, 8]). In a second scenario, assuming that we perform record linkage on an arbitrary number of databases, the features used as variables in the downstream tasks can be present in more than one database. Exact error propagation for this general case has not been explored in the literature to the best of our knowledge. With this in mind, we investigate the effect of record linkage on general downstream tasks, where we assume that features are present in all databases and we allow for duplication within a database (Section 4).

In this paper, we are primarily concerned with downstream tasks that fall under the umbrella of generalized linear models. We fit these models with a Bayesian specification, with Gaussian prior distributions for the parameters [6, 7], with predictors centered and scaled to be weakly informative. More specifically, we assume the following: Y denotes the response vector, X denotes the $n \times p$ dimensional covariate matrix, N_p denotes the p dimensional Normal distribution, and I_p denotes the p dimensional identity matrix. The hyperparameters b_i and a are left as the weakly informative defaults in stan [20]. The linear regression model is specified as

$$Y|\beta, \sigma, X \sim N_p(X\beta, \sigma^2 I_p)$$

 $\beta_i|b_i \stackrel{ind}{\sim} N(0, b_i)$
 $\sigma|a \sim \text{Exponential}(a),$ (3.1)

and the logistic model specified as

$$Y_i | \boldsymbol{\beta}, \boldsymbol{X}_i \overset{ind}{\sim} \text{Bernoulli} \left(\frac{\exp\{\boldsymbol{X}_i^T \boldsymbol{\beta}\}}{1 + \exp\{\boldsymbol{X}_i^T \boldsymbol{\beta}\}} \right)$$

$$\beta_i | b_i \overset{ind}{\sim} \text{N}(0, b_i). \tag{3.2}$$

Remark: We consider the above models as a proof of concept for our prototyping methods, given that our goal in this paper is to focus on the prototyping methods and their performance, rather than fine tuning regression models.

4 Experiments

We consider two experiments to assess the performance of our proposed methods of the most representative record. We describe the simulated data (Section 4.1), define performance evaluation metrics (Section 4.1),

 $^{^{2}}$ Note that a record with PP weight above 0.5 has a posterior probability greater than 0.5 of being chosen as a prototype and should be included in the final data set.

and present our findings. First, we present a "best case scenario" in which the record linkage was able to perfectly capture the linkage structure in the data (Section 4.2). Second, we present a more realistic scenario in which we assess the performance of each of the three data set creation methods after performing record linkage using model (2.2). This scenario is split into two cases – the most general case, in which all variables in the downstream task are subject to record linkage error and a more common case in which only the explanatory variables are subject to record linkage error.

4.1 Data and Evaluation Metrics

In this section, we describe the simulated data used in all empirical studies and metrics used to evaluate our proposed methodology.

4.1.1 Data

In this section, we describe the simulated data used throughout our experiments. For all empirical studies, we simulated three data sets with different levels of noise in the relationship between predictors and response variables through Gaussian noise, where $\sigma_{\epsilon}=1,2,5$. The data sets contain a total of 500 records, 30% duplication, and the maximum number of duplicates of each record is 5. Each data set contains the following features: first name, last name, birth date, sex, education level, income (in 1000s), bp, and high_bp. The bp and high_bp variables were generated with a known relationship to sex and income; and our goal is to assess how the fitted models are altered based on the representative data set passed from the linkage model.

We generated the three data sets with 500 records using the GeCO tool [26], where each data set consists of the following features: first name, last name, and birth date. We use optical character recognition, keyboard edit, phonetic edit, and common misspellings to distort the name features in 150 and duplicate records in each data set.³ Next, we add the following features to the original data: sex, education level, income (in 1000s), and bp. In order to add sex to the data set, we used the babynames package [30] in R [17] and matched each first name and year of birth with the closest match in US baby names from 1880-2015. If a name did not appear in the look-up data set, we randomly assigned male or female with equal probability. To sample education level, we used United States (US) Census information on educational attainment in the US for individuals over the age of 25 [28] to get conditional distributions by age, group, and sex. For the income variable, we sampled from a Gaussian distribution with $\sigma = 5$ and mean taken from the median earnings by educational attainment and gender from the US Bureau of Labor Statistics [27]. For the systolic by variable, we created a model with two main assumptions – men have higher bp than women and income is inversely related to bp. The model we generated from is

$$bp = 160 + 10\mathbb{I}(sex = \text{``M''}) - income + 0.5income * \mathbb{I}(sex = \text{``M''}) + \epsilon$$

where $\epsilon \sim \text{Normal}(0, \sigma_{\epsilon}^2)$ and $\sigma_{\epsilon} = 1, 2, 5$, for the three different noise levels that correspond to the three data sets. Additionally, we generated a high-bp variable, which is binary according to the following.

$$\begin{split} & \text{high_bp}_i|\text{income}_i, \text{sex}_i \overset{iid}{\sim} \text{Bernoulli}(p_i) \\ & \log \frac{p_i}{1-p_i} = 30 + 10\mathbb{I}(\text{sex}_i = \text{``M"}) - \text{income}_i + 0.5\text{income}_i * \mathbb{I}(\text{sex}_i = \text{``M"}) + \epsilon \end{split}$$

where, again, $\epsilon \sim \text{Normal}(0, \sigma_{\epsilon}^2)$ and $\sigma_{\epsilon} = 1, 2, 5$, for the three different noise levels that correspond to the three data sets.

Additionally, we generated three sets of test records, of 500 records each following the same data generation mechanism. These records are used to evaluate the performance of the downstream task after using the prototyping methods in Section 3.1.

Given the three data sets generated, we add distortion to the duplicate records for the training data. Recall, first name and last name have already been distorted. To distort the other features, we choose three out of the five remaining fields (birth date, sex, education level, income, and bp) to distort and then alter them according to the rules below.

³We allow for up to 5 duplicates for each record and each attribute can be distorted up to two times for each record.

Column	Distortion rule
birth date	Add random noise to the date according to Normal(0, 25) distribution
sex	Sample male or female with equal probability
education level	Sample from the existing education levels with equal probability
income	Sample from the existing income values with equal probability
blood pressure	Sample from the existing blood pressure values with equal probability

Table 3: Rules for adding distortion to the duplicates according to each column type.

4.1.2 Evaluation Metrics

In this section, we describe the metrics used to evaluate our proposed methodology. We first describe standard record linkage metrics utilized and then describe our evaluation metrics for prototyping (choosing a representative data set) and the downstream task.

We evaluate the quality of record linkage performance by comparing the clustering provided in each MCMC iteration to the true clustering used to generate the data and computing precision and recall [25, 4]. Precision and recall are defined as

$$\begin{aligned} \text{Precision} &= \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \\ \text{Recall} &= \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}, \end{aligned}$$

where the true positives is the number of record pairs correctly linked, true negatives is the number of record pairs predicted correctly to not be linked, false positives is the number of records pairs predicted incorrectly to be linked, and false negatives is the number of record pairs predicted incorrectly to be not linked.

Next, we evaluate the quality of our proposed prototyping methods for choosing a representative data set using two ways. First, we assess the distributional closeness of the representative data set to the true records. The distributional closeness of the representative data sets to the true records is useful because one of the benefits of using a two-stage approach to record linkage and downstream analyses is the ability to perform multiple analyses with the same data set. As such, downstream performance of representative records may be dependent on the type of downstream task that is being performed. In order to assess the distributional closeness of the representative data sets to the truth, we use an empirical Kullback-Leibler (KL) divergence metric [29, 19]. Let $\hat{F}_{rep}(x)$ and $\hat{F}_{true}(x)$ be the empirical distribution functions for the representative data set and true data set, respectively (with continuous variables transformed to categorical using a histogram approach with statistically equivalent data-dependent bins, as in [29]). The empirical KL divergence metric we use is then defined as

$$D_{KL}(\hat{F}_{rep}||\hat{F}_{true}) = \sum_{\boldsymbol{x}} \hat{F}_{rep}(\boldsymbol{x}) \log \left(\frac{\hat{F}_{rep}(\boldsymbol{x})}{\hat{F}_{true}(\boldsymbol{x})} \right).$$

Second, we assess the performance of the downstream task using multiple approaches. One approach is determining if the credible intervals for the coefficients in the model contain the true values. Another way is evaluating how well the downstream task performs on the test data set. Based on the downstream task this involves using the fitted model to predict the outcome variable on the test data set. If the model based on a representative data set performs similar to the model based on the true records, then we can say the method of creating the representative data set is performing well. More specifically, for regression tasks, we evaluate models using the mean squared errors (MSE) and for binary classification and record linkage tasks, we evaluate models using the predicted probability of being classified as having high blood pressure as well as precision and recall for prediction of the test data. In all cases, we compare the predictions from the true model to the actual true values present in the data set.

4.2 Prototyping with Known Clusters

As a baseline, we first generate the representative records from the known clusters of the three simulated data sets. This situation is a "best case scenario", where the record linkage model can link the records

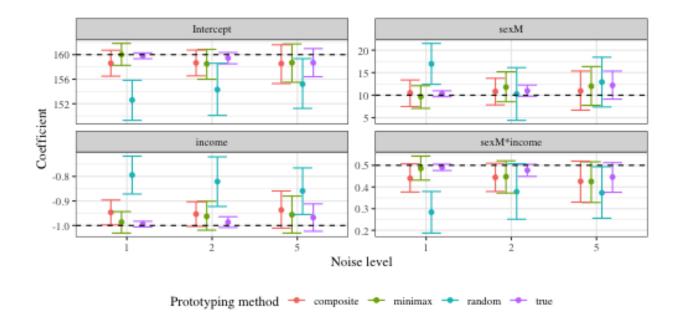


Figure 1: 95% credible intervals of regression coefficients for the models fit from three methods of prototyping (random, minimax, and composite) and from the true records for noise levels of $\sigma_{\epsilon} = 1, 2, 5$. The horizontal dashed lines show the true values (known from simulation) of each coefficient. For each coefficient and noise level, minimax and composite prototyping perform closely to the model resulting from the true records and captures the true model parameters, whereas the random prototyping method shows greater differences.

perfectly based on their true underlying relationship.

Table 4 shows the empirical KL divergence $(D_{KL}(\hat{F}_{rep}||\hat{F}_{true}))$, see Section 4.1 for details) for minimax, composite, and random prototyping as compared to the true record values. Note that a smaller value of the empirical KL divergence indicates that the variables to be used in the downstream task (bp, high bp, income, and sex) from the representative data set are closer in distribution to the true record values. For all noise levels, the minimax prototyping method shows the best performance, indicating that it should outperform the other methods for downstream tasks using these variables.

σ_{ϵ}	Minimax	Composite	Random
1	0.01	0.04	0.02
2	0.01	0.06	0.04
5	0.02	0.1	0.06

Table 4: Empirical KL divergence $(D_{KL}(\hat{F}_{rep}||\hat{F}_{true}))$ for the result of applying three prototyping methods for selecting a representative data set – composite, minimax prototyping, and random prototyping – as compared to true record values for three levels of noise in the relationship. Minimax prototyping outperforms the other methods for all noise levels.

We use the representative records resulting from the prototyping methods to then fit linear and logistic regression models with Gaussian distribution priors for the parameters, as specified in Section 3.2. In this experiment, where the record clusters are known without the need for a record linkage step, we compare the random prototyping (random), minimax prototyping (minimax), and the composite methods. Performance is assessed by how well each model predicts the dependent variable in the test set compared to the model fit to the true records and whether the true parameters are captured in the credible intervals of the model coefficients.

The linear regression model was fit using the known form of relationship between bp, income, and sex,

 $E[bp|income, sex] = \beta_0 + \beta_1 \mathbb{I}(sex = "M") + \beta_2 income + \beta_3 income * \mathbb{I}(sex = "M")$ with the representative data sets resulting from the three methods (random, minimax, and composite) and the true records, for comparison. Figure 1 shows the posterior distributions of the model coefficients for each proposed method for selecting representative records (three noise levels). For each coefficient and noise level, minimax and composite prototyping perform closely to the model resulting from the true records and captures the true model parameters, whereas the random prototyping method shows greater differences.

Table 5 confirms these results, showing the mean squared errors (MSE) for the linear regression models of bp on income and sex for each method of selecting a representative data set as well as the MSE from selecting the true record for three levels of noise in the relationship. Again, minimax prototyping outperforms the other methods, in that it is the most similar to the true MSE for all noise levels. The superior performance of minimax prototyping here is intuitive. The number of false records selected by uniform random prototyping is 31, 41, and 39, for the three noise levels, while the number for minimax prototyping method is 10, 10, and 10. This aligns with the results from inspecting the empirical KL divergence metric. Due to the distance function chosen, minimax prototyping always outperforms uniform random selection. The weighted average composite records will alter all records with a duplicate, necessarily. In this case, 51 records are affected. This makes the composite prototyping method particularly sensitive to outliers from noisy copies of data, whereas the minimax prototyping method is more robust. Thus, the regression, where the representative record is chosen by minimax prototyping, outperforms the other two methods for the linear regression case.

σ_{ϵ}	true	minimax	composite	random
1	1.07	1.31	2.31	5.57
2	4.19	4.44	4.98	11.65
5	25.2	25.89	26.67	27.7

Table 5: Mean squared error (MSE) for linear regression models of bp on income and sex for three methods of selecting a representative data set – composite, minimax prototyping, and random prototyping – as well from selecting the true record for three levels of noise in the relationship. Minimax prototyping outperforms the other methods for all noise levels.

The logistic regression model was fit using the known (from simulation) relationship between high bp, income, and sex, $\log \frac{p}{1-p} = \beta_0 + \beta_1 \mathbb{I}(\text{sex} = \text{"M"}) + \beta_2 \text{income} + \beta_3 \text{income} * \mathbb{I}(\text{sex} = \text{"M"})$, where p = p(high bp|income,sex), with the data sets created from the three methods as well as the true records. Figure 2 illustrates the results from the logistic model of high bp status on income and sex. For records that have high bp, a good model will have a predicted value on the scale of the response close to 1, and for those without high bp a value close to 0. This is seen across all noise levels for the model fit to the true data. This is also true for the minimax method, however the random and composite prototyping methods result in response values closer to 0.5 for both types of records. This is more pronounced for the composite method as the noise increases, emphasizing the sensitivity of the composite prototyping method to noisy data. This indicates that the minimax prototyping method illustrates superior performance to the random and composite methods in the logistic regression task as well for the "best case scenario" where record linkage is perfect.

4.3 Prototyping with Record Linkage

In this section, we compare the effect of the prototyping methods in the downstream tasks after linking the records using model $(2.2)^4$. We explore the performance of prototyping methods from Section 3.1 in two potential data scenarios. The first scenario is where all the variables used in the downstream task (explanatory and response) are available in all data sources, and thus susceptible to error from the record linkage process. In the second scenario, we consider a more realistic problem, where only the the explanatory variables are available from multiple sources, and thus, the response is not distorted between data sets. Before describing these two situations, we first describe the record linkage method used and how representative data sets are constructed from this method.

⁴The model was fit using the package eber [1]; diagnostics are given in Appendix A.

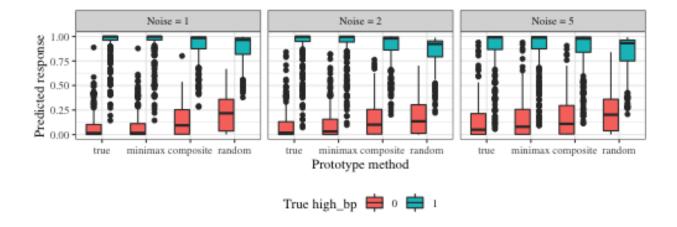


Figure 2: Distribution of predicted values from the logistic regression model fit from the result of three prototyping methods (random, minimax, and composite) as well as from the true data for three different noise levels of $\sigma_{\epsilon} = 1, 2, 5$. The blue boxplots are the distributions of predicted response for records with high bp and the red for those without high bp. For records that have high bp, a good model will have a predicted value on the scale of the response close to 1, and for those without high bp a value close to 0.

4.3.1 Record Linkage and the Representative Data Sets

As mentioned, in this section, we detail the record linkage process and the relationship between record linkage and prototyping. In both scenarios to follow, we first perform record linkage using Model (2.2) with the features first name, last name, and birth date. Given posterior samples of the linkage structure Λ , we can utilize either the distribution of the linkage structure or a point estimate of an optimal single linkage given by the shared most probable maximal matching sets (MPMMS) introduced in [21, 24] (see Appendix B for details). The precision and recall associated with the MPMMS are 0.9 and 0.97. We expect the performance of the regression analyses to be affected by this linkage error for all the representative record selection methods. We explore variations of the proposed methods by involving the pairwise posterior linkage probabilities in the representative record selection. These probabilities are computed from Λ as the number of times a pair of records was assigned to the same cluster over the number of MCMC iterations. In particular, we utilize the pairwise linkage probabilities to construct weights for each record within a cluster to sample a representative record at random (pairwise random) and to generate composite records (pairwise composite) according to these weights. For record selection with minimax prototyping, we use the complement of the linkage probabilities as a distance measure between record pairs (pairwise minimax).

As an alternative to selecting or constructing a representative record set from the MPMMS clusters, we have proposed posterior prototyping (PP) weights to perform a weighted regression analysis that naturally propagates the linkage error. In the experimental results, the record linkage performed reasonably well at each iteration, with mean precision and recall of 0.86 and 0.93, respectively (see Appendix A for full trace plots of precision and recall for each chain). As explained in Section 3.1, for each MCMC iteration from model (2.2), we obtain the most representative records using minimax prototyping and then compute the probability of each record being selected over all MCMC iterations to create the posterior probability weights (PP weights). Recall, the PP weights are used in two ways to propagate errors to the downstream task – as weights for each record in the regression and as a thresholding metric to create a representative data set. Figure 3 displays a minimax PP weights distribution for the true and duplicated records in the data set with $\sigma_{\epsilon} = 1$ as an example. Note that the true records consistently have higher PP weights and the proportion of duplicated records with high weights is relatively low.

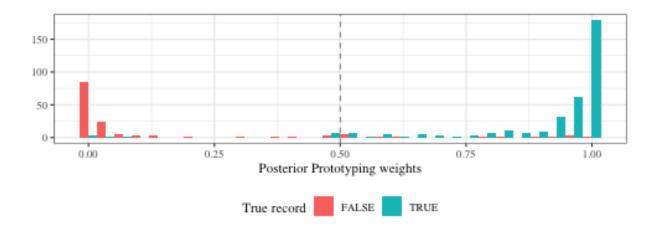


Figure 3: The distribution of PP weights as generated from posterior draws of Λ colored by if they are true or duplicated records in the original data set. The dotted vertical line shows the threshold value of 0.5. The true records have consistently higher PP weights and the proportion of duplicated records with high weights is relatively low.

4.3.2 Errors in All Downstream Variables

Before assessing the fit of downstream models on representative data sets, we first examine the closeness of the representative data sets to the true records in two ways. Figure 4 (top panel) displays a summary of the number of false records selected by all the random and minimax prototyping methods using 100 data sets of representative records. Observe that the number of false records ranges between 5 and 55 approximately, and that the PP threshold prototype based on the minimax with variable-level distance consistently outperforms the other prototyping methods. The bottom panel shows the empirical KL divergence for 100 data sets of representative data sets compared to the true records. In this case, the posterior prototyping methods (weighted and threshold) show the closest distributions to the truth. Additionally, in these methods (and the minimax prototyping), there is not a wide range in the empirical KL divergence values, indicating that the distance function is performing an adequate job of discerning records such that there are not many ties resulting.

The range of values present in the minimax, pairwise minimax, and posterior prototyping (PP weighted and PP thresh) results in Figure 4 is solely due to the randomness resulting from ties. In particular, the large variation for the pairwise minimax method suggests that the complement of the pairwise posterior probabilities is not a very discriminatory distance function. This is also evidenced by the similar performance of the random and pairwise random methods. Due to the fact that we only consider pairwise posterior probabilities within the optimal MPMMS clustering and the linkage variables have a high discriminatory power, the resulting pairwise probabilities are high and very similar for all pairs of records within a cluster.

Next, we examine the performance of the prototyping methods through the lens of the results from each downstream task. Table 6 displays the MSE for linear regression models for three levels of noise in the relationship of bp on income and sex for all our proposed methods for selecting a representative data set — PP threshold, PP weighted, minimax, composite, pairwise posterior composite, random, pairwise posterior minimax and random prototyping — as well as results with the true records. The PP threshold and minimax methods display superior performance for all noise levels, with PP weighted as close behind. The composite method also shows decent performance on the test data set, but will not be robust to outliers in the variables used for the downstream task, and so should be used with caution if there is high distortion in the data.

Figures 5 and 6 shows the results from the logistic model of high bp status on income and sex. The first shows the distribution of predicted values from the fitted model. Again, for records that have high bp, a good model will have a predicted value on the scale of the response close to 1, and for those without high bp a value close to 0. This is seen across all noise levels for the model fit to the true data. The methods based on posterior prototyping and minimax show similar results to the truth. The composite methods

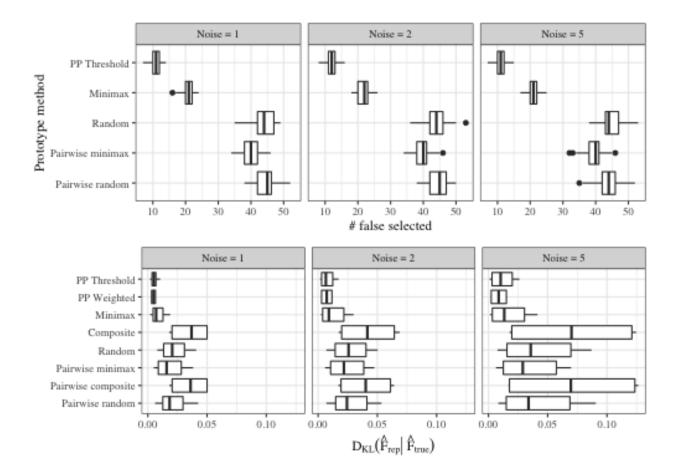


Figure 4: Top: The distribution of number of false records selected by each prototyping method by performing each method 100 times. Minimax prototyping based on the distance function results in the least number of incorrect records as prototypes. Bottom: Empirical KL divergence for each prototyping method compared to the true records, performed 100 times. The posterior prototyping methods and minimax yield results that are closest to the truth, in all noise levels.

Method	Noise = 1	Noise $= 2$	Noise $= 5$
True	1.07	4.19	25.2
PP Threshold	$1.84 \ (0.24)$	4.73(0.18)	$25.69 \ (0.15)$
PP Weighted	1.94(0.03)	4.75(0.03)	25.95 (0.04)
Minimax	1.95(0.31)	$4.55 \ (0.17)$	25.94(0.32)
Composite	2.97(0.12)	5.52(0.11)	27.18(0.2)
Random	6.16(2.11)	8.47(1.83)	29.86(2.06)
Pairwise minimax	5.57(1.52)	8.07(1.78)	29.09(1.51)
Pairwise composite	2.96(0.12)	5.48(0.11)	27.15(0.2)
Pairwise random	6.05 (1.84)	8.81(2.13)	29.74 (2.07)

Table 6: Mean squared error (MSE) for linear regression models of bp on income and sex for all proposed methods for selecting a representative data set as well as the true data for levels of noise $\sigma=1,2,5$ in the relationship between the predictors and the response variable. For probabilistic methods, the mean of 100 generated data sets is shown with the standard deviation in parentheses. The PP threshold method outperforms the other methods for all noise levels, with PP weighted regression and minimax protoyping second best. The performances of the pairwise minimax and both variants of the random method are considerably inferior compared to the other approaches.

show decreasing performance as the noise increases, which is to be expected, as it will be very sensitive to outliers. This conclusion is confirmed in the precision and recall for each model as predictions on a test data set (Figure 6). The minimax and posterior prototyping methods show the most consistent performance, especially as the noise in the data set increases. Composite methods also show good results when the noise level is low, but performance deteriorates as noise increases.

While their performance is similar in all metrics explored, the posterior prototyping methods have the advantage over the minimax method of propagating error from the record linkage task through the downstream task. This is evident in the credible intervals of downstream model coefficients that follow from the posterior prototyping methods, versus those based on point estimates (MPMMS) from record linkage. This is shown in Figure 7, where the 95% credible intervals for the downstream task coefficient estimates for linear regression based on the minimax and posterior prototyping methods are compared for each noise level. The minimax method intervals do not contain the true values of the coefficients (horizontal black lines) more often then the posterior prototyping methods, as seen in red. The PP weighted method produces very wide intervals to account for the record linkage error, and the PP threshold method achieves more coverage than the minimax method with narrower credible intervals than the PP weighted method.

4.3.3 Errors in Explanatory Variables Only

Here we look at the same assessments for closeness of distributions and downstream performance for a less general, but more realistic scenario. We are interested in how all prototyping methods perform when the response variable for the downstream task is potentially not available in multiple data sets, and thus not susceptible to record linkage error. This is a common assumption in the record linkage literature for developing downstream methods [5, 32, 9], as well as in many applications [16, 12, 15].

Figure 8 shows the empirical KL divergence for the representative data sets compared to the true record values, performed 100 times. The posterior prototyping methods and minimax yield results that are closest to the truth, in all noise levels, matching the results from the more general case. The downstream performance assessments for linear and logistic regression are presented in Table 7 and Figures 9 and 10, respectively. Compared to the case when the response variable is subject to record linkage error, again the minimax and the PP threshold methods appear to have the best performance. In the logistic regression case, Figure 10 shows minimax and posterior prototyping methods to have by far the most consistent performance, very close to the performance of the true model. Meaning, these methods perform almost as well as the benchmark, and are not subject to the same randomness as the other, inferior, methods.

In closing, for the case where the response variable is not repeated across multiple data sets, and thus not subject to record linkage error, the posterior prototyping methods and the minimax prototyping method show high performance across all noise levels. The difference between these methods being that the uncertainty

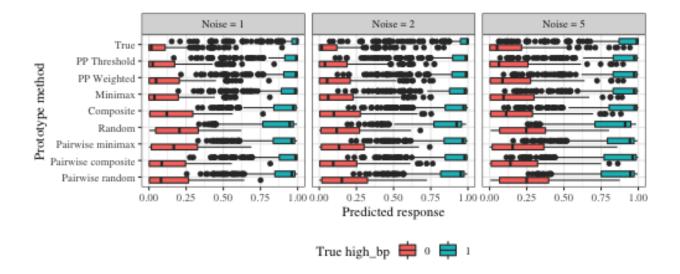


Figure 5: Distribution of predicted values from the logistic regression model fit from the result of all prototyping methods as well as from the true data for three different noise levels of $\sigma_{\epsilon} = 1, 2, 5$. The blue boxplots are the distributions of predicted response for records with high bp and the red for those without high bp. For records that have high bp, a good model will have a predicted value on the scale of the response close to 1, and for those without high bp a value close to 0. Posterior prototyping and minimax methods show the best performance, while the composite method's performance deteriorates as the data becomes more noisy.

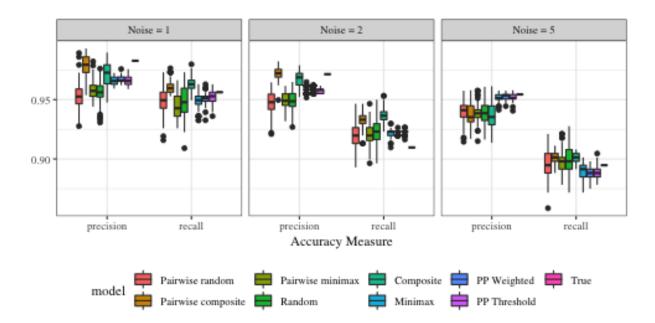


Figure 6: Precision and recall from the logistic regression model fit from the result of all prototyping methods as well as from the true data for three different noise levels of $\sigma_{\epsilon} = 1, 2, 5$ as prediction on a test data set. The minimax and posterior prototyping methods show the most consistent performance, especially as the noise in the dataset increases. Composite methods also show good results when the noise level is low, but performance deteriorates as noise increases.

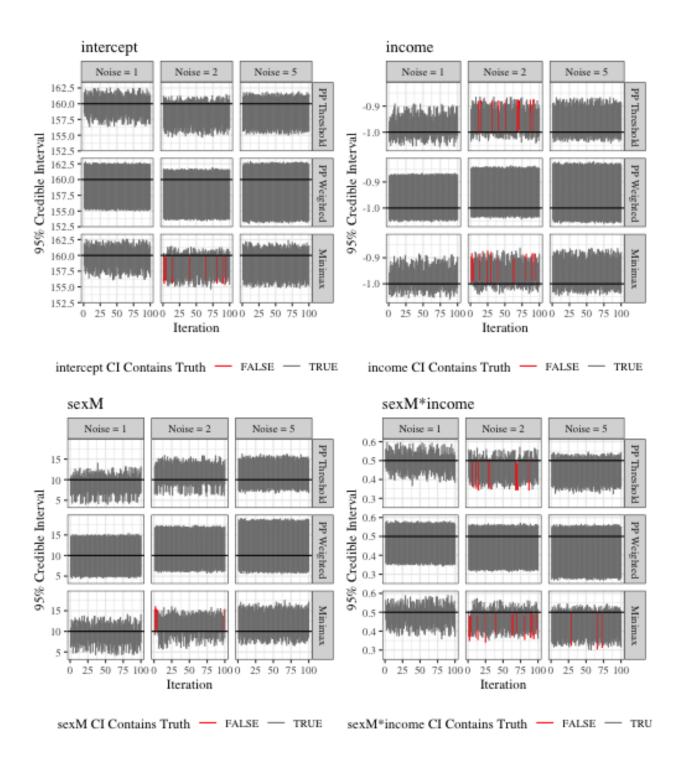


Figure 7: Comparison of 95% credible intervals for the downstream task coefficient estimates for linear regression based on the minimax and posterior prototyping methods for each noise level, repeated 100 times. The minimax method intervals do not contain the true values of the coefficients (horizontal black lines) more often then the posterior prototyping methods. The PP weighted method produces very wide intervals to account for the record linkage error.

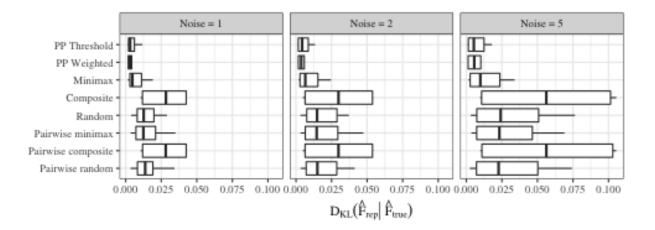


Figure 8: Empirical KL divergence for each prototyping method compared to the true records, performed 100 times, for the case where the response variable is not subject to record linkage error. The posterior prototyping methods and minimax yield results that are closest to the truth, in all noise levels, matching the results from the more general case.

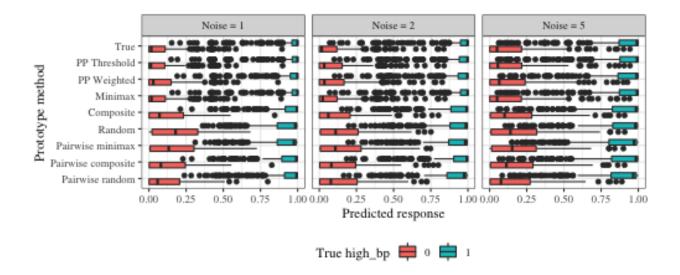


Figure 9: Distribution of predicted values from the logistic regression model fit from the result of all prototyping methods as well as from the true data for three different noise levels of $\sigma_{\epsilon} = 1, 2, 5$ for the scenario when the response variable is not repeated across data sets. The blue boxplots are the distributions of predicted response for records with high bp and the red for those without high bp. Recall, for records that have high bp, a good model will have a predicted value on the scale of the response close to 1, and for those without high bp a value close to 0. Posterior prototyping and minimax methods show the best performance, while the composite method's performance deteriorates as the data becomes more noisy.

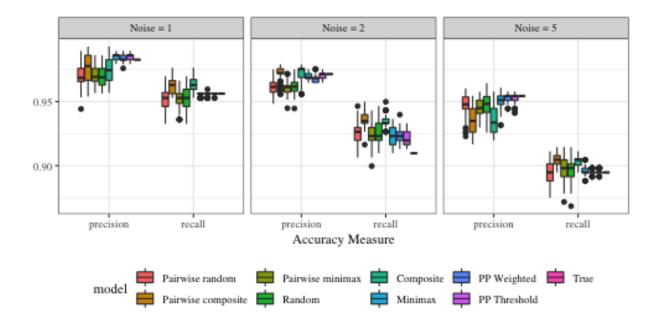


Figure 10: Precision and recall from the logistic regression model fit from the result of all prototyping methods as well as from the true data for three different noise levels of $\sigma_{\epsilon} = 1, 2, 5$ as prediction on a test data set for the scenario when the response variable is not repeated across data sets. The minimax and posterior prototyping methods show the most consistent performance, especially as the noise in the dataset increases. These methods also show performance close to the true model, meaning they are close to the benchmark set.

Method	Noise = 1	Noise $= 2$	Noise $= 5$
True	1.07	4.19	25.2
PP Threshold	1.14(0.05)	4.25 (0.06)	25.27(0.09)
PP Weighted	1.18(0.01)	4.31(0.01)	25.32(0.02)
Minimax	$1.13\ (0.04)$	4.25 (0.06)	25.24 (0.1)
Composite	1.4(0.09)	4.54(0.1)	25.75(0.11)
Random	3.7(1.36)	6.65(1.45)	27.59(1.56)
Pairwise minimax	3.42(0.92)	6.54 (0.89)	27.48 (0.94)
Pairwise composite	1.41(0.1)	4.55(0.12)	25.74(0.12)
Pairwise random	3.76(1.48)	6.87(1.8)	27.65 (1.24)

Table 7: Mean squared error (MSE) for linear regression models of bp on income and sex for all proposed methods for selecting a representative data set as well as the true data for levels of noise $\sigma=1,2,5$ in the relationship between the predictors and the response variable. For probabilistic methods, the mean of 100 generated data sets is shown with the standard deviation in parentheses. The PP threshold method outperforms the other methods for all noise levels, with PP weighted regression and minimax protoyping second best. The performances of the pairwise minimax and both variants of the random method are considerably inferior compared to the other approaches.

from record linkage is propagated to the downstream task in the posterior prototyping methods, whereas only a point estimate is used for the minimax method. This leads to more realistic credible intervals for the coefficients in the downstream task. Something else of note, is that the PP weighted regression method shows less variability due to ties because the standard deviation of values over all 100 repeats of the process is very small for all noise levels. This indicates that the PP Weighted method is a very stable method, even if it is not achieving the very best performance as evaluated via the results of the downstream task.

5 Prototyping in Practice, Discussion, and Future Directions

We have proposed many prototyping methods to create a representative data set for downstream tasks. In addition, we have evaluated our methods in many simulated situations. In this section, we first describe how a practitioner could utilize this in practice and then provide a general discussion of our main contributions to the literature.

5.1 Practitioner Guidance

In this section, we provide practical practitioner guidance for using uniform, minimax, composite, or posterior prototyping.

Random prototyping is a probabilistic method that is the simplest prototyping method one could use. In essence, this is the benchmark for the practitioner to compare to given its simplicity and ease of computation. In general, we expect the other three proposed prototype methods to outperform the random prototyping method. There are some exceptions for when the random prototype method will behave well, and we describe one such situation below.

Turning to minimax prototyping and posterior prototyping (PP weighted and PP threshold), there are inherent assumptions about the properties of the data that allow these methods to perform well. All of these methods can only show improved performance over the random prototype if the distance function is a good discriminator between records. If the distance function is not able to discern the difference between records, the each method will result in many tied values. Thus, random selection will perform roughly the same as these other methods. In this paper, (see Section 3.1), we have chosen a record distance function that performs discrimination between the records quite well, and we quantify this by the empirical KL divergence. On the other hand, suppose we could only calculate a distance based on the following record linkage variables: first name, last name, birthdate. Then the minimax prototype method with column-wise distance would result in 49.02% of the clusters to be resolved being tied, versus 27.45%, 27.45%, and 27.45% for using the full data

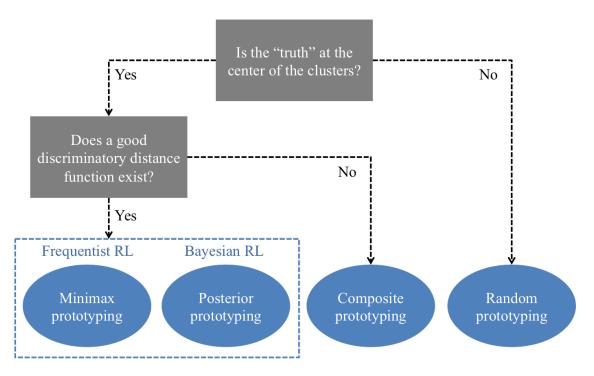


Figure 11: A potential decision making process for determining which prototyping method to use.

set with noise levels 1, 2, and 5, respectively. Therefore, this improved ability to discern differences among the records is the basis of performance improvements for methods based on minimax prototyping.

The composite prototyping method (in addition to the minimax-based methods), both rely on the assumption that the "truth" lies at the center in p-dimensional space of the records that have been clustered together. If this is not the case, then these methods will not perform well. One example of this is linking longitudinal records, where the most recent record is more likely to be accurate. In this case, perhaps random selection with a very informative distribution would out-perform these centroid-based methods.

In general, we recommend the use of posterior prototyping when three conditions are met – Bayesian record linkage is used, the true records lie central to their clusters, and a good discriminatory distance function is available. When Bayesian record linkage is not used, then we recommend minimax prototyping and when a discriminatory distance function is not available we recommend composite prototyping. Finally, in the case where there is good reason to believe that the true records do not lie central to their clusters, then we are left with only random prototyping. In this case, the use of a very informative distribution for records within cluster may be helpful.

Due to the fact that each application of prototyping could call for many decisions to be made by a potential user, Figure 5.1 shows a decision making process for determining which prototyping method to use. In general, we advocate for the use of Bayesian methods, due to their ability to naturally incorporate uncertainty into the downstream task. However, in the event that this is not possible, ther are many practical questions the user will want to consider to choose a prototyping method. First, the practitioner will want to consider the following question: "Is it reasonable to assume the true records lie in the center of the clustered records?" If the answer is yes, then centroid-based methods (like minimax-based or composite) are reasonable choices. Second a practitioner will want to consider the following question: "Does there exist a good discriminatory distance function?" This is crucial to the success of minimax-based prototyping, and so, if the answer is "No", then we recommend composite methods to be used. Finally, the practitioner must recognize that the choice of the record linkage method can impact the choice in the prototyping method. For example, if a Bayesian record linage method is used, then we recommend posterior prototyping due to the fact that it respects the Bayesian paradigm and also allow for natural error propagation. In addition, the results from Section 4 also suggest that this method performs well in practice based upon our performance metrics.

5.2 Discussion and Future Work

We have proposed four prototyping methods to select the most representative records from data with duplicated records, which can be passed to the downstream task. In addition, we have explored how error propagates from the linkage process into the downstream task.

Minimax prototyping renders the best results in our baseline simulation studies without the use of record linkage and provides a building block to construct record weights from the posterior that can be used to propagate the linkage error to the downstream task, allowing for the methodology to be fully Bayesian throughout. While the posterior prototyping methods add computational cost, they provide the added benefit of allowing for natural error propagation into the downstream task. In addition, the computation burden depends largely on the size and percentage of duplicates in the data. While it is computationally more expensive than the alternatives, in most record linkage tasks the cluster sizes remain small including a large number of singletons clusters [2]. Thus, the pairwise distance computation for minimax prototyping in each MCMC iteration adds a relatively low computational burden to the overall process. This suggests that our proposed methodology has practical potential and should be explored further in both simulated and real data scenarios.

A final benefit of our proposed methodology for prototyping is its generality. In fact, prototyping can be used with any record linkage method and any downstream task. Many open and future areas of research include understanding the trade offs regarding changing the choice of the record linkage model under simulated and real data. In addition, it is of interest to consider general downstream tasks, such as general linear models, small area estimation, and general classification problems.

References

- [1] Australian Bureau of Statistics. *eber: Empirical Bayes Entity Resolution*, To appear. R package version 0.1.
- [2] Brenda Betancourt, Giacomo Zanella, Jeffrey W Miller, Hanna Wallach, Abbas Zaidi, and Rebecca C Steorts. Flexible models for microclustering with application to entity resolution. In Advances in Neural Information Processing Systems, pages 1417–1425, 2016.
- [3] Jacob Bien and Robert Tibshirani. Hierarchical clustering with prototypes via minimax linkage. *Journal of the American Statistical Association*, 106(495):1075–1084, 2011.
- [4] Peter Christen. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer-Verlag Berlin Heidelberg, 2012.
- [5] Nicole M. Dalzell and Jerome P. Reiter. Regression modeling and file matching using possibly erroneous matching variables. *Journal of Computational and Graphical Statistics*, 0(0):1–11, 2018.
- [6] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.*, 2(4):1360–1383, 12 2008.
- [7] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. Bayesian data analysis. Chapman and Hall/CRC, 2013.
- [8] Harvey Goldstein, Katie Harron, and Angie Wade. The analysis of record-linked data using multiple imputation with data value priors. *Statistics in Medicine*, 31(28):3481–3493, 2012.
- [9] Roee Gutman, Christopher C. Afendulis, and Alan M. Zaslavsky. A bayesian procedure for file linking to analyze end-of-life medical costs. *Journal of the American Statistical Association*, 108(501):34–47, 2013.
- [10] M. H. P. Hof and A. H. Zwinderman. Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. *Statistics in Medicine*, 31(30):4231–4242, 2012.
- [11] Michel H. Hof, Anita C. Ravelli, and Aeilko H. Zwinderman To. A probabilistic record linkage model for survival data. *Journal of the American Statistical Association*, 112(520):1504–1515, 2017.

- [12] Rune Jacobsen, Erik Bostofte, Gerda Engholm, Johnni Hansen, Niels E Skakkebæk, and Henrik Møller. Fertility and offspring sex ratio of men who develop testicular cancer: a record linkage study. *Human Reproduction*, 15(9):1958–1961, 2000.
- [13] Gunky Kim and Raymond Chambers. Regression analysis under incomplete linkage. *Comput. Stat. Data Anal.*, 56(9):2756–2770, September 2012.
- [14] P Lahiri and Michael D Larsen. Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469):222–230, 2005.
- [15] Glyn Lewis and Andy Sloggett. Suicide, deprivation, and unemployment: record linkage study. *Bmj*, 317(7168):1283–1286, 1998.
- [16] Stephen C Newman and Roger C Bland. Mortality in a cohort of patients with schizophrenia: a record linkage study. *The Canadian Journal of Psychiatry*, 36(4):239–245, 1991.
- [17] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2018.
- [18] M. Sadinle and S. E. Fienberg. A generalized fellegi-sunter framework for multiple record linkage with application to homicide record systems. *Journal of the American Statistical Association*, 108(502):385–397, 2013.
- [19] Jorge Silva and Shrikanth Narayanan. Universal consistency of data-driven partitions for divergence estimation. In *Information Theory*, 2007. ISIT 2007. IEEE International Symposium on, pages 2021– 2025. IEEE, 2007.
- [20] Stan Development Team. rstanarm: Bayesian applied regression modeling via Stan., 2016. R package version 2.13.1.
- [21] Rebecca Steorts, Rob Hall, and Stephen Fienberg. Smered: A bayesian approach to graphical record linkage and de-duplication. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 922–930, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.
- [22] Rebecca C Steorts. Entity resolution with empirically motivated priors. *Bayesian Analysis*, 10(4):849–875, 2015.
- [23] Rebecca C. Steorts, Mattew Barnes, and Willie Neiswanger. Performance Bounds for Graphical Record Linkage. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 298–306, Fort Lauderdale, FL, USA, April 2017. PMLR.
- [24] Rebecca C. Steorts, Rob Hall, and Stephen E Fienberg. A bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516):1660–1672, 2016.
- [25] Rebecca C. Steorts, Samuel L. Ventura, Mauricio Sadinle, and Stephen E. Fienberg. A comparison of blocking methods for record linkage. In Josep Domingo-Ferrer, editor, *Privacy in Statistical Databases*, pages 253–268, Cham, 2014. Springer International Publishing.
- [26] Khoi-Nguyen Tran, Dinusha Vatsalan, and Peter Christen. Geco: an online personal data generator and corruptor. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2473–2476. ACM, 2013.
- [27] United States Bureau of Labor Statistics. Median weekly earnings by educational attainment in 2014. https://www.bls.gov/opub/ted/2015/median-weekly-earnings-by-education-gender-race-and-ethnicity-in-2014.htm, January 2015.
- [28] United States Census Bureau. Educational attainment in the united states: 2017. https://census.gov/data/tables/2017/demo/education-attainment/cps-detailed-tables.html, December 2017.

- [29] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9):3064–3074, 2005.
- [30] Hadley Wickham. babynames: US Baby Names 1880-2015, 2017. R package version 0.3.0.
- [31] W. E. Winkler. Overview of record linkage and current research directions. Technical report, U.S. Bureau of the Census Statistical Research Division, 2006.
- [32] Joan Heck Wortman and Jerome P Reiter. Simultaneous record linkage and causal inference with propensity score subclassification. *Statistics in Medicine*, 2018.

A Record Linkage MCMC Diagnostic Plots

In this section, we provide diagnostic plots for the record linkage analysis. We look at three functions of Λ — number of unique entities, precision, and recall — to assess convergence and mixing of the chain.

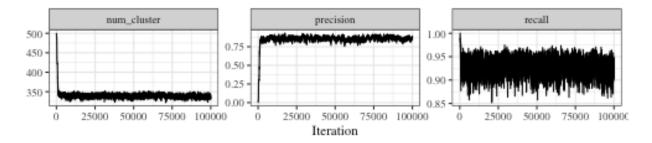


Figure 12: Trace plots of three functions of Λ — number of unique entities, precision, and recall — to check for convergence of the chain.

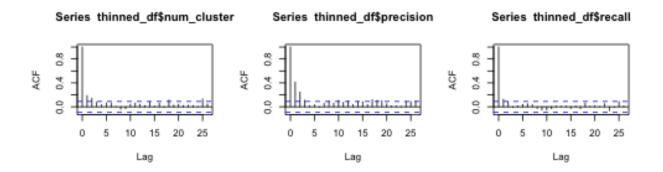


Figure 13: Autocorrelation functions plotted for multiple lags of three functions of Λ — number of unique entities, precision, and recall — after burn-in and thinning, to check for mixing performance.

B Most Probable Maximal Matching Set (MPMMS)

To obtain a point estimate of the partition of the data into clusters from the linkage structure Λ , [22] defines a maximal matching set (MMS) as the set containing all the records associated with the same latent entity. For a particular record, its most probable MMS (MPMMS) is defined as the set with the highest posterior probability of being an MMS. However, it is possible to have incongruent most probable maximal matching sets for record pairs. For example, record A may be in the most probable maximal matching set of record B, but record B may not be in the most probable maximal matching set of record A. To address this issue, the optimal clustering is obtained by linking records in the same shared MPMMS that corresponds to the most probable MMS for each its members. The final point estimate of the linkage structure respects the transitive property of matched record pairs within each cluster.