Simultaneous Edit-Imputation for Continuous Microdata

Journal:	*Journal of the American Statistical Association*
Manuscript ID:	JASA-A&CS-2014-0107.R2
Manuscript Type:	Article – Applications & Case Studies
Keywords:	Bayesian, Economic, Editing, Missing, Survey

Simultaneous Edit-Imputation for Continuous Microdata

Abstract

Many statistical organizations collect data that are expected to satisfy linear constraints; as examples, component variables should sum to total variables, and ratios of pairs of variables should be bounded by expert-specified constants. When reported data violate constraints, organizations identify and replace values potentially in error in a process known as edit-imputation. To date, most approaches separate the error localization and imputation steps, typically using optimization methods to identify the variables to change followed by hot deck imputation. We present an approach that fully integrates editing and imputation for continuous microdata under linear constraints. Our approach relies on a Bayesian hierarchical model that includes (i) a flexible joint probability model for the underlying true values of the data with support only on the set of values that satisfy all editing constraints, (ii) a model for latent indicators of the variables that are in error, and (iii) a model for the reported responses for variables in error. We illustrate the potential advantages of the Bayesian editing approach over existing approaches using simulation studies. We apply the model to edit faulty data from the 2007 U.S. Census of Manufactures. Supplementary materials for this article are available online.

Keywords: Bayesian; Economic; Editing; Missing; Mixture; Survey

1. INTRODUCTION

Many statistical organizations collect data that include faulty values, i.e., implausible or inconsistent data. Left uncorrected, faulty values can hamper the organization's analysis and interpretation of the data, and can undermine the public's confidence in the quality of files disseminated by the organization. Thus, many organizations spend substantial resources correcting faulty data in a process known as edit-imputation. For example, Granquist and Kovar (1997) estimated that national statistical agencies spend 40% of total budgets for business surveys on edit-imputation processes, and Norberg (2009) reported that Statistics Sweden allocated 32.6% of total costs in business surveys to edit-imputation processes. Edit-imputation is also a key concept in the total survey error paradigm employed by many statistical agencies (Groves and Lyberg 2010; Biemer 2010).

Organizations may be able to perform edit-imputation for some records by re-contacting respondents to ascertain correct values, or by leveraging other data sources like administrative records with (possibly) correct values. Often, however, these options are too expensive or not available for all records with faulty values. In such cases, organizations typically rely on automatic editing, in which mathematical algorithms identify and correct faulty values with minimal human intervention (De Waal and Coutinho 2005; Pannekoek et al. 2013).

Most automatic editing systems in use today run in two steps, an error localization step in which some set of each faulty record's variables is determined to be incorrect, and an imputation step in which those values are replaced with plausibly correct values (De Waal and Coutinho 2005; De Waal et al. 2011). To understand the need for error localization, consider three variables required to satisfy $x_1 + x_2 = x_3$. If this relationship does not hold for a record in the reported data, the failure of the equality potentially could result from any one of the three variables being incorrect, from any pair being incorrect, or from all three being incorrect; the editing system must choose among these possibilities. Most editing systems use an approach to error localization suggested by Fellegi and Holt (1976), namely to find a solution that changes the minimum number of fields (variables) to make the faulty record satisfy all constraints (e.g., see Greenberg and Surdi 1984; Kovar et al. 1988; De Waal 1996; Draper and Winkler 1997; Pannekoek and De Waal 2005; De Jonge and Van der Loo 2011, 2014). The subsequent imputation step usually is a variant of hot deck

imputation, although model-based approaches also have been proposed (e.g., Raghunathan et al. 2001; Tempelman 2007; Parker and Schenker 2007; Kim et al. 2014).

While widely used, Fellegi and Holt approaches to edit-imputation—henceforth abbreviated as F-H approaches—have potential drawbacks. First, they do not fully utilize the information in the data in the error localization process. To illustrate, consider an example where the variables include sex, pregnancy status, and age in years. If a record is reported as a pregnant male who is 40 years old, it seems more plausible to change status to not pregnant than to change sex to female, because of the association between age and pregnancy status. The minimum number of fields criterion admits changing either one of sex or status. The organization would select among these two solutions based on some heuristic, e.g., change the variable that is more likely to have errors according to the organization's experience in other contexts. Second, the process of error localization inherently has uncertainty—the organization generally cannot be certain that a F-H (or any) error localization has identified the exact locations of errors—that is ignored by specifying a single error localization; hence, analyses of the data underestimate uncertainty. For example, there is a chance that the person is a pregnant 40 year old woman rather than a not pregnant 40 year old man, and inferences should reflect that possibility (as well as other feasible variations) as increased uncertainty. Third, for complicated systems of constraints involving many variables, the original F-H approach can be computationally time consuming to implement in practice (Garcia 2002; De Waal et al. 2011). Thus, many F-H edit systems use alternative optimization algorithms to reduce computational burdens, for example, cutting plane methods (Garfinkel et al. 1988; Riera-Ledesma and Salazar-González 2007), branch-and-bound algorithms (De Waal 1996; Barcaroli and Venturi 1997; De Waal and Quere 2003; Pannekoek and De Waal 2005), or algorithms to find suboptimal solutions (Schiopu-Kratina and Kovar 1989; Draper and Winkler 1997; Winkler and Chen 2002; Garcia 2002).

In this article, we propose an integrated approach to edit-imputation that addresses some of the shortcomings of the F-H paradigm. Our approach relies on a Bayesian hierarchical model that includes (i) a flexible joint probability model for the underlying true values of the data with support only on the set of values that satisfy all editing constraints,

(ii) a model for latent indicators of the variables that are in error, and (iii) a model for the reported responses for variables in error. We present this hierarchical model for data with only continuous variables—similar ideas could be used for other data settings—and consider editing constraints in the form of both inequalities and equalities on the relationships among the variables. Since this model fully integrates the error localization and imputation steps, it takes advantage of relationships in the data when identifying faulty values, appropriately incorporates uncertainty over the space of the latent error indicators, and by design imputes corrected values that are guaranteed to satisfy all constraints. The MCMC algorithm for estimating the model generates corrected datasets as by-products; these can be released by the organization as public use files and analyzed using multiple imputation combining rules (Rubin 1987).

Our approach is similar in spirit to other approaches to dealing with measurement error. Ghosh-Dastidar and Schafer (2003) use a hierarchical model like ours, with a multivariate normal distribution for underlying true values, independent binomial distributions for error indicators, and independent normal distributions for reporting errors. Little and Smith (1987) use Mahalanobis distances to select faulty values, replacing outliers with imputations based on a multivariate normal distribution. These approaches are not designed to handle complex systems of equality and inequality constraints, as they can generate imputations that violate constraints. They also are based on stronger distributional assumptions than the underlying mixture model we use.

The remainder of the article is organized as follows. In Section 2, we review automatic editing, especially focusing on F-H approaches. In Section 3, we present the Bayesian hierarchical model for edit-imputation. In Section 4, we report results of simulation studies comparing the proposed method with F-H and other edit-imputation approaches. In Section 5, we apply the model to edit a set of faulty records from the 2007 U.S. Census of Manufactures. In Section 6, we conclude with a discussion of future research directions.

2. REVIEW OF AUTOMATIC EDITING

For $i = 1, \ldots, n$, let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ be the true values of p variables for data subject i, and let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})$ be the corresponding reported values. Any difference between $\boldsymbol{x}_i$

and $\boldsymbol{y}_i$ arises from some source(s) of measurement or response errors, e.g., resulting from respondent mistakes, interviewer or mode effects, and unclear wording of survey questionnaires (Groves 1989). For $i = 1, \ldots, n$ and $j = 1, \ldots, p$, let $s_{ij} = 0$ when $y_{ij} = x_{ij}$ and $s_{ij} = 1$ otherwise (i.e., when variable j is in error), and let $\boldsymbol{s}_i = (s_{i1}, \ldots, s_{ip})$. We also set $s_{ij} = 1$ for any y_{ij} that is missing in the reported data. Following usual nomenclature (Garfinkel et al. 1988; Kovar et al. 1988; United Nations 2006; Norberg 2009), we refer to any variable with $s_{ij} = 1$ as a flagged item or, equivalently, a field to impute.

Many national statistical agencies and survey organizations have developed sets of logical conditions, called edit rules or simply edits, to determine if $\boldsymbol{y}_i$ is faulty ($\sum_j s_{ij} > 0$) or not ($\sum_j s_{ij} = 0$). Edit rules typically are developed by subject matter experts at the agency, designed to identify data that are structurally impossible, for example individuals having more years of work experience than years of life, or practically implausible, for example a statistics professor making a billion dollars in salary per year. For continuous data, typical edit rules include *range restrictions* of the form $L_j \leq y_{ij} \leq U_j$, where L_j and U_j are variable-specific constants. For example, in economic data many variables are required to take on non-negative values; any reported record with negative values is deemed faulty. Variables also can be subject to *ratio edits* of the form $L_{j,j'} \leq y_{ij}/y_{ij'} \leq U_{j,j'}$, where $L_{j,j'}$ and $U_{j,j'}$ are constants bounding the ratio of variable j to variable j'. For example, an agency may decide that every business establishment's total annual wages divided by its number of employees should be greater than \$1,000 and less than \$1,000,000; any record with reported wages and employee size outside this plausible region is deemed faulty. Finally, continuous variables can be subject to *balance edits* in which one variable must equal the sum of others. For example, the total number of employees equals the sum of the number of production workers and the number of other employees. The entire collection of range restrictions, ratio edits, and balance edits defines a feasible region comprising all potential records passing the edits, which we denote by $\mathcal{X}$. For exemplary systems of edit rules for continuous data, see Winkler and Draper (1996), Garcia and Goodwin (2002), and Thompson et al. (2004). We note that data also may contain conditional edits, for example if $y_{ij} > 0$ then $y_{ij'} > 0$; we do not deal with conditional edits here.

Automatic editing systems based on F-H methods specify $\boldsymbol{s}_i$ using the minimum fields

to impute (MFI) criterion. Specifically, for each record Fellegi and Holt (1976) find the solution $\boldsymbol{s}_i$ that minimizes $\sum_j s_{ij}$ yet still can result in edited values that lie in $\mathcal{X}$; all missing fields are forced to have $s_{ij} = 1$. The key insight of Fellegi and Holt (1976) is that one can use the complete set of explicit edits, i.e., those written by the subject matter experts, and implied edits, i.e., extra conditions logically derived from explicit edits, in order to efficiently find the optimal solution of $\boldsymbol{s}$ under the MFI criterion. In practice, many systems find the $\boldsymbol{s}_i$ that minimizes $\sum_j w_j s_{ij}$, where w_j (called the reliability weight of variable j) is relatively larger for variables pre-determined by the agency as less likely to be in error. This is referred to as the minimum weighted fields to impute (MWFI) criterion (Garfinkel et al. 1986).

We now illustrate how implied edits are useful for finding MWFI (or MFI) solutions. Suppose that an agency specifies two explicit edits, $x_1 \leq x_2$ and $x_2 \leq x_3$ and reliability weights $(w_1, w_2, w_3) = (1, 2, 3)$. Suppose that $(y_{i1}, y_{i2}, y_{i3}) = (6, 4, 2)$, which fails both explicit edits. Without considering implied edits, the optimal MWFI solution at first glance appears to be $(s_{i1}, s_{i2}, s_{i3}) = (0, 1, 0)$, i.e., change x_2, because each failed edit involves x_2. However, $\boldsymbol{s}_i = (0, 1, 0)$ is not feasible; no value of x_2 makes $(6, x_2, 2)$ satisfy both explicit edits. Once we add the implied edit $x_1 \leq x_3$, it becomes immediately apparent that the only feasible solutions include $(1, 1, 0)$, $(1, 0, 1)$, $(0, 1, 1)$, and of course $(1, 1, 1)$. The MWFI criterion selects $\boldsymbol{s}_i = (1, 1, 0)$ to minimize $\sum_j w_j s_{ij}$. We note that the implied edits serve to help identify MFI solutions efficiently. They are not needed to check if a proposed solution is feasible, which can be done by determining whether it passes or fails the explicit edits.

Arguably, underlying the use of the MFI or MWFI criterion is a philosophy that the organization should use as much of the respondent-supplied data as possible. While an understandable position, it can lead to potentially unrealistic distributions of corrected values. Consider Figure 1, which depicts a setting with range restrictions and a ratio edit for two variables. For the faulty case in the lower right corner of the figure, the MFI criterion would change only one of the two variables. With either choice, the corrected value after the imputation step is in the extreme tail of the empirical bivariate distribution of the edit-passing records. It may well be that this is reasonable; on the other hand, it may not. Other variables on the file could suggest that neither value is likely to be correct.

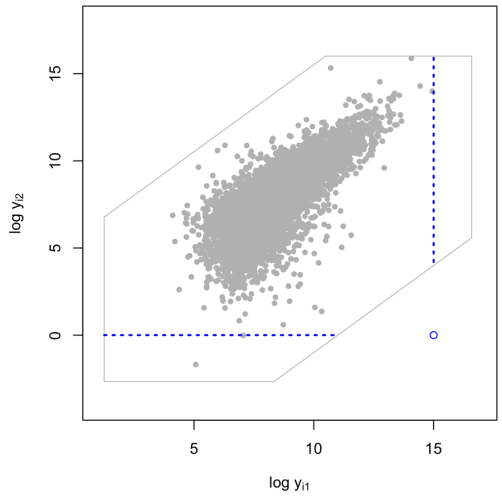

Figure 1: Illustrative example of the MFI criterion leading potentially to implausible corrected values. The hexagon represents the feasible region formed by ratio edits and range restrictions on two variables in the log scale. The solid dots in the hexagon are edit-passing records, and the open dot outside the hexagon is a faulty record. The dotted lines in the feasible region represent the support of the imputed values for the faulty record under a MFI criterion. The resulting imputations will be in the extreme tails of the distribution of the edit-passing records.

In the scenario of Figure 1, it is advantageous to allow both fields to change as informed by relationships in the data. This is done in the Bayesian editing model we now describe.

3. BAYESIAN EDITING MODEL

Prior to describing the Bayesian editing model, we introduce some notation for balance edits. Suppose that there are q distinct balance edits, $\{B_1, \ldots, B_q\}$, in a system of edits. For all records i subject to B_l, the variables involved in B_l comprise a single total variable with true and reported values (x_{iT_l}, y_{iT_l}), and two or more component variables. For notational simplicity, henceforth we use B_l to represent both the l-th balance edit and the indexes of the component variables involved in that edit, with true and reported values $\{x_{ij} : j \in B_l\}$ and $\{y_{ij} : j \in B_l\}$. Here, $\sum_{j \in B_l} x_{ij} = x_{iT_l}$ always, but it is not necessarily the case that $\sum_{j \in B_l} y_{ij} = y_{iT_l}$. Define the set of all-but-total variables by $NT = \{1, \ldots, p\} \setminus \{T_l : l =$

$1, \ldots, q\}$, which includes all component variables and variables not involved in any balance edits. For each i, let $\boldsymbol{x}_{i,NT} = \{x_{ij} : j \in NT\}$ and $\boldsymbol{y}_{i,NT} = \{y_{ij} : j \in NT\}$ respectively be the true values and the reported values of variables in NT. With nested balance edits, for example, $x_{i1} = x_{i2} + x_{i3}$ while $x_{i3} = x_{i4} + x_{i5}$, we include in NT variables that are not totals, so that we can re-express the balance edits as sums of components. For example, we make the two balance edits $x_{i1} = x_{i2} + x_{i4} + x_{i5}$ and $x_{i3} = x_{i4} + x_{i5}$, and let $NT = (2, 4, 5)$.

We assume that the feasible region $\mathcal{X}$ is defined by a combination of balance edits, ratio edits, and range restrictions. We assume that $\mathcal{X}$ is bounded and not empty; that is, the system of edits is consistent and has solutions. For convenience, we define $\mathcal{D}$ to be the convex polytope defined by only ratio edits and range restrictions. Thus, we can characterize $\mathcal{X}$ as the space in which all $\boldsymbol{x}_i \in \mathcal{D}$ and $\sum_{j \in B_l} x_{ij} = x_{iT_l}$ for $l = 1, \ldots, q$. We note that it is prudent for agencies to check the feasibility of $\mathcal{D}$ before edit-imputation.

We consider scenarios where any $\boldsymbol{y}_i$ that passes all edits is treated as a true value, i.e., $s_{ij} = 0$ and $x_{ij} = y_{ij}$ for all j. We also set $s_{ij} = 0$ for any other value known to be correct (e.g., from manual edits), and we fix $s_{ij} = 1$ when y_{ij} is missing or violates range restrictions. For the remaining cases, we leave s_{ij} unspecified. By treating $y_{ij} = x_{ij}$ for all j when $\boldsymbol{y}_i$ passes all edits, we mimic the actions of automatic editing systems in use by many national statistical agencies including, for example, the Census Bureau, Statistics Canada, and Statistics Netherlands (De Waal et al. 2011).

To facilitate model specification, it is convenient to categorize records based on if and how they violate the edit constraints. Specifically, we define $A_i = 0$ when record i satisfies all edits ($\sum_j s_{ij} = 0$); $A_i = 1$ when record i fails at least one balance edit and passes all inequality constraints; $A_i = 2$ when record i passes all balance edits but fails at least one inequality constraint; and $A_i = 3$ when record i fails at least one balance edit and at least one inequality constraint. Each A_i is completely determined given $\boldsymbol{y}_i$ and the edit rules, but not necessarily given $\boldsymbol{s}_i$.

We first present the model in a general form, then present a specification of the model relevant for editing the Census of Manufactures (CM) data.

3.1 General Form of Bayesian Editing Model

We use a Bayesian hierarchical model with three levels (plus prior distributions) including a model for $\boldsymbol{x}_i$, a model for $(\boldsymbol{s}_i, A_i)$ given $\boldsymbol{x}_i$, and a model for $\boldsymbol{y}_i$ given $(\boldsymbol{x}_i, \boldsymbol{s}_i, A_i)$. The model for $\boldsymbol{x}_i$ can be any multivariate continuous distribution with support only on $\mathcal{X}$. Letting $\boldsymbol{\theta}$ be the parameters of this model, we write the model as

$$f(\boldsymbol{x}_i|\boldsymbol{\theta}) = f\left(\boldsymbol{x}_{i,NT}|\boldsymbol{\theta}\right) \prod_{l=1}^{q} \delta\left(x_{iT_l} - \sum_{j \in B_l} x_{ij}\right) I\left[\boldsymbol{x}_i \in \mathcal{D}\right] \tag{1}$$

for $i = 1, \ldots, n$, where $f(\boldsymbol{x}_{i,NT}|\boldsymbol{\theta})$ is a $(p-q)$-dimensional joint density for the all-but-total variables, $\delta(\cdot)$ is the Dirac delta function with the point mass at zero, and $I[\cdot] = 1$ if the statement inside the brackets is true and $I[\cdot] = 0$ otherwise. This model restricts totals to equal the sum of their components, and restricts all variables to lie within $\mathcal{D}$.

For the error indicators $(\boldsymbol{s}_i, A_i)$, a generic form of the model is $f(\boldsymbol{s}_i, A_i \mid \boldsymbol{x}_i, \psi_s)$, where ψ_s are parameters assumed to be distinct from $\boldsymbol{\theta}$. This generic form allows errors potentially to depend on $\boldsymbol{x}_i$; for example, records with small values of some x_{ij} are more likely to have $s_{ij} = 1$ (or simply to have $A_i > 0$) than records with large values of x_{ij}. In this case, one possible model is a logistic regression of s_{ij} on variables in $\boldsymbol{x}_i$, in which case ψ_{sj} includes the regression coefficients. More simply, one could favor certain combinations of $\boldsymbol{s}_i$ over others within categories A_i. For example, one could assume $Pr(s_{ij} = 1 \mid A_i \neq 0, \boldsymbol{x}_i, \psi_{sj}) = \psi_{sj}$, giving higher prior probability to changing some variables over others, e.g., when total variables are deemed more reliable than component variables.

For the reported values $\boldsymbol{y}_i$, we use measurement error models for variables with $s_{ij} = 1$ and set $y_{ij} = x_{ij}$ whenever $s_{ij} = 0$. Given $(\boldsymbol{x}_i, \boldsymbol{s}_i, A_i)$, we partition $\boldsymbol{y}_i = (\boldsymbol{y}_i^{UF}, \boldsymbol{y}_i^{F})$, where $\boldsymbol{y}_i^{UF} = \{y_{ij} : s_{ij} = 0, j = 1, \ldots, p\}$ are correctly reported (not flagged) values and $\boldsymbol{y}_i^{F} = \{y_{ij} : s_{ij} = 1, j = 1, \ldots, p\}$ are incorrectly reported (flagged) values. Either set may be empty. We partition $\boldsymbol{x}_i = (\boldsymbol{x}_i^{UF}, \boldsymbol{x}_i^{F})$ corresponding to the same $\boldsymbol{s}_i$. These partitions generally are not fixed for faulty records; rather, they are redefined each time we draw a new value of $\boldsymbol{s}_i$ in the MCMC algorithm described in Section 3.3. We write the reporting model with parameters ψ_y as

$$f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{s}_i, A_i, \psi_y) = f\left(\boldsymbol{y}_i^{F} \mid \boldsymbol{x}_i, A_i, \psi_y\right) \delta\left(\boldsymbol{y}_i^{UF} - \boldsymbol{x}_i^{UF}\right).$$

Illustrative reporting models include, for example, $y_{ij}^F \sim N(x_{ij}^F, \sigma_j^2)$ and, in the absence of any information about reporting error, $f(y_{ij}^F) \propto 1$. The formulation allows the reporting model to differ with A_i. It can be sensible to use different models for y_{ij}^F due to reporting error and y_{ij}^F due to missing data, e.g., a normal distribution for reporting error and a uniform distribution for missing data.

This construction allows the distribution of the edited values to depend explicitly on the observed data. To see this, consider the conditional distribution of $(\boldsymbol{x}_i, \boldsymbol{s}_i)$, assuming all variables are subject to edits. Ignoring parameters for notational simplicity, we have

$$f(\boldsymbol{x}_i, \boldsymbol{s}_i \mid \boldsymbol{y}_i, A_i) \propto f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{s}_i, A_i) f(\boldsymbol{s}_i, A_i \mid \boldsymbol{x}_i) f(\boldsymbol{x}_i). \tag{2}$$

Thus, the model favors edit-imputations that (i) are not unlikely under the posited model for reporting error, (ii) are not unlikely under the posited model for error indicators, and (iii) are not unlikely under the posited model for the underlying data. For example, if we assume $f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{s}_i, A_i) \propto 1$ and $f(\boldsymbol{s}_i, A_i \mid \boldsymbol{x}_i) \propto 1$, as we do in Section 3.2, the model favors changing reported values that have low likelihood according the model for $\boldsymbol{x}_i$, such as outliers or combinations of variables that do not accord with the correlation structure in the data.

Because the model is modular, it can be easily integrated with other editing strategies. For example, agency experts can identify values clearly in error, such as outliers known from external information or previous history to be implausible, as recommended in Kozak (2005) and Banff Support Team (2007), and set their corresponding $s_{ij} = 1$ before running the Bayesian editing procedure. This way, the agency ensures replacing these values with imputations. As another example, agencies can use automatic editing procedures to detect typing errors and sign errors (Scholtus 2009, 2011; Van der Loo et al. 2011; Pannekoek et al. 2013). Before running the Bayesian editing procedure, the agency can correct such errors if true values can be inferred, or otherwise set the corresponding $s_{ij} = 1$.

Some analysts may want to use only records that pass all edits, or perhaps records that pass all edits for the subset of variables relevant to a particular analysis. These scenarios are akin to complete-cases and available-cases analyses, respectively, in the standard missing data literature (Little and Rubin 2002). To characterize when such analyses can be valid, we define *faulty at random* (FAR) mechanisms that satisfy the following criteria. Let $\mathcal{C}_i$

index the set of variables known to have $s_{ij} = 0$, with corresponding values $\boldsymbol{y}_i^{\mathcal{C}_i} = \boldsymbol{x}_i^{\mathcal{C}_i}$ and $\boldsymbol{s}_i^{\mathcal{C}_i} = 0$. Let $\mathcal{E}_i$ be the set of remaining values for record i, i.e., all cases with missing or possibly erroneous y_{ij}, with corresponding values $(\boldsymbol{x}_i^{\mathcal{E}_i}, \boldsymbol{s}_i^{\mathcal{E}_i}, \boldsymbol{y}_i^{\mathcal{E}_i})$. A mechanism is FAR if for each record the likelihood for $(\boldsymbol{s}_i^{\mathcal{E}_i}, \boldsymbol{y}_i^{\mathcal{E}_i}, A_i)$ is conditionally independent of $\boldsymbol{x}_i^{\mathcal{E}_i}$, i.e.,

$$f(\boldsymbol{s}_i^{\mathcal{E}_i}, \boldsymbol{y}_i^{\mathcal{E}_i}, A_i \mid \boldsymbol{x}_i^{\mathcal{E}_i}, \boldsymbol{x}_i^{\mathcal{C}_i}, \boldsymbol{s}_i^{\mathcal{C}_i}, \boldsymbol{y}_i^{\mathcal{C}_i}, \psi_s, \psi_y, \boldsymbol{\theta}) = f(\boldsymbol{s}_i^{\mathcal{E}_i}, \boldsymbol{y}_i^{\mathcal{E}_i}, A_i \mid \boldsymbol{x}_i^{\mathcal{C}_i}, \boldsymbol{s}_i^{\mathcal{C}_i}, \boldsymbol{y}_i^{\mathcal{C}_i}, \psi_s, \psi_y, \boldsymbol{\theta}).$$

We note that the conditioning need only be on any two variables in $(\boldsymbol{x}_i^{\mathcal{C}_i}, \boldsymbol{s}_i^{\mathcal{C}_i}, \boldsymbol{y}_i^{\mathcal{C}_i})$, but we leave all three to emphasize the distinction between $\mathcal{C}_i$ and $\mathcal{E}_i$. FAR implies that any two records with the same values of $(\boldsymbol{x}_i^{\mathcal{C}_i}, \boldsymbol{s}_i^{\mathcal{C}_i}, \boldsymbol{y}_i^{\mathcal{C}_i})$ have the same distribution of $\boldsymbol{x}_i^{\mathcal{E}_i}$, as the reported values $\boldsymbol{y}_i^{\mathcal{E}_i}$ and indicators of errors $\boldsymbol{s}_i^{\mathcal{E}_i}$ do not depend on $\boldsymbol{x}_i^{\mathcal{E}_i}$. Embedded in FAR is the condition that x_{ij} is missing at random (Rubin 1976) for any record with $s_{ij} = 1$ because y_{ij} is not reported. Assuming (ψ_s, ψ_y) is distinct from $\boldsymbol{\theta}$, analyses of complete or available cases are valid under FAR mechanisms provided that the variables in $\boldsymbol{x}_i^{\mathcal{C}_i}$ are accounted for in the analysis, e.g., they are included as predictors in regression models. The proof of this statement is similar to the proof for standard ignorable missing data mechanisms (Little and Rubin 2002), in that the likelihood for $\boldsymbol{\theta}$ can be made free of $\boldsymbol{x}_i^{\mathcal{E}_i}$ under these conditions.

3.2 Model Specification for the CM Data

We now specify particular true data, error indicator, and reporting models for the CM data. For a graphical summary of all model components, see Appendix A of the supplementary material.

True Data Model. In the CM and other databases with economic variables, the joint distribution of $\boldsymbol{x}$ has complex features not easily captured by standard multivariate distributions, such as non-Gaussian tails and nonlinear dependencies. Thus, we prefer a flexible underlying model for $\boldsymbol{x}$; for continuous multivariate data, one such model is the mixture of multivariate normal distributions (MacEachern and Müller 1998). However, we also must restrict the support of $\boldsymbol{x}$ to $\mathcal{X}$, which includes balance equations and inequality constraints. To do so, we extend the approach of Kim et al. (2014), who propose a finite mixture of multivariate normal distributions for fixed $\boldsymbol{s}_i$ with support on arbitrary $\mathcal{D}$ but no balance

edits. As in Kim et al. (2014), we work with the natural logarithms of all non-zero values, setting $\log(0) = \log(\omega)$ where ω is a pre-determined, small number. Although not necessary, using logarithms facilitates computation and prior specification (Little and Smith 1987).

Specifically, we suppose that each individual belongs to exactly one of K mixture components. For all i, let $z_i \in \{1, \ldots, K\}$ be a latent indicator of mixture component membership, and let $Pr(z_i = k) = \pi_k$ where $\sum_k \pi_k = 1$. Let $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ be respectively the mean vector and the covariance matrix of $\log \boldsymbol{x}_{i,NT}$ in the k-th mixture component; let $\mu = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ and $\Sigma = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$; let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$; and, let $\boldsymbol{\theta} = (\mu, \Sigma, \boldsymbol{\pi})$. Thus, for (1) we use

$$\boldsymbol{x}_i|\mu, \Sigma, z_i \propto \mathrm{N}\left(\log \boldsymbol{x}_{i,NT}|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}\right) \prod_{l=1}^{q} \delta\left(x_{iT_l} - \sum_{j \in B_l} x_{ij}\right) I\left[\boldsymbol{x}_i \in \mathcal{D}\right], \tag{3}$$

$$z_i \sim \text{Categorical}(\pi_1, \ldots, \pi_K). \tag{4}$$

For μ and Σ, we use conjugate prior distributions (Lavine and West 1992). For $k = 1, \ldots, K$ and $j = 1, \ldots, p - q$, we have

$$\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k \sim \mathrm{N}(\boldsymbol{\mu}_0, h_0^{-1}\boldsymbol{\Sigma}_k), \tag{5}$$

$$\boldsymbol{\Sigma}_k \sim \text{InverseWishart}(\zeta_0, \boldsymbol{\Phi}), \quad \boldsymbol{\Phi} = diag(\Phi_1, \ldots, \Phi_{p-q}), \tag{6}$$

$$\Phi_j \sim \text{Gamma}(a_\Phi, b_\Phi), \tag{7}$$

where $\text{Gamma}(a, b)$ denotes the Gamma distribution with mean a/b. In the simulations and application to the CM data, we set $a_\Phi = b_\Phi = 0.25$ to put substantial prior mass on modest-sized variances; $\zeta_0 = (p - q) + 1$ to ensure proper distributions; $\boldsymbol{\mu}_0$ equal to the mean of $\log \boldsymbol{y}$ from edit-passing records only; and $h_0 = 5$.

For $\boldsymbol{\pi}$, we use a finite Dirichlet process (Ishwaran and James 2001), using the stick-breaking representation of Sethuraman (1994). We have

$$\pi_k = v_k \prod_{g<k}(1 - v_g) \quad \text{for } k = 1, \ldots, K, \tag{8}$$

$$v_k \sim \text{Beta}(1, \alpha) \quad \text{for } k = 1, \ldots, K - 1; \quad v_K = 1, \tag{9}$$

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha). \tag{10}$$

Following Dunson and Xing (2009), we set $a_\alpha = b_\alpha = 0.25$, which represents a vague specification for the Gamma prior distribution. Sensitivity analyses in the simulations and CM analysis show no practical differences for other sensible choices of $(a_\Phi, b_\Phi, a_\alpha, b_\alpha, h_0)$, which mirrors a finding of Kim et al. (2014).

We recommend setting K to be large enough to capture patterns in the data, but not so large so as to create many components with no observed data. Analysts can examine the posterior distribution of the number of unique values of z_i among the n observed cases during the MCMC fitting to diagnose if K is large enough. Significant posterior mass at a number of classes equal to K suggests that K be increased. For the simulations and CM application, we use $K = 50$.

Error Indicator Model. For $i = 1, \ldots, n$, we assume that

$$f(A_i \mid \boldsymbol{x}_i, \psi_s) \propto 1, \quad f(\boldsymbol{s}_i \mid A_i = 0, \boldsymbol{x}_i, \psi_s) = \delta(\boldsymbol{s}_i = \mathbf{0}), \quad f(\boldsymbol{s}_i \mid A_i \neq 0, \boldsymbol{x}_i, \psi_s) \propto 1. \quad (11)$$

This implies that, for any record i, *a priori* all candidate $\boldsymbol{s}_i$ that can result in feasible solutions (with $s_{ij} = 0$ for all variables not subject to edits) for that record are equally likely. The uniform distribution is computationally convenient and allows the observed data to drive the error localization. We note that, provided A_i does not depend on x_{ij} involved in edit failures, the exact nature of the prior distribution on A_i in (11) is irrelevant since A_i is determined by $\boldsymbol{y}_i$ and the edit rules.

Reporting Model. When specifying models for $\boldsymbol{y}_i$, we divide records into three types: passing all edits ($A_i = 0$), failing some balance edits ($A_i = 1, A_i = 3$), and passing all balance edits but failing some inequality constraints ($A_i = 2$). As an example of a record with $A_i = 2$, consider a respondent that reports true values for some component variables, faulty values for other component variables as determined by ratio edits or range restrictions, and the total variable as the sum of the component values. We combine $A_i = 1$ and $A_i = 3$ because, as we describe below, the support of $\boldsymbol{y}_i$ in both cases can be defined by the same mathematical expression, whereas the support of $\boldsymbol{y}_i$ when $A_i = 2$ has a different form.

For records with $A_i = 0$, we follow the CM editing practice and set $f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{s}_i, A_i, \psi_y) = \prod_{j=1}^{p} \delta\left(y_{ij} - x_{ij}\right)$. For records with failed edits we use uniform distributions to represent vague knowledge about the nature of reporting errors. In discussions with Census Bureau

personnel about editing processes for the CM, we were not able to identify more specific measurement error models. With uniform reporting distributions, given $(\boldsymbol{x}_i^{UF}, \boldsymbol{s}_i, A_i)$ the value of $\boldsymbol{y}_i^F$ does not provide any information about $\boldsymbol{x}_i^F$ and so can be disregarded, even for missing y_{ij}.

When $A_i \in \{1, 3\}$, we have

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{s}_i, A_i, \psi_y) = \prod_{\{j:s_{ij}=1, j\in NT\}} \text{Unif}\left(\log y_{ij} \in [\log \tilde{L}_j, \log \tilde{U}_j]\right) \times \prod_{\{j:s_{ij}=0\}} \delta\left(y_{ij} - x_{ij}\right) \prod_{l\in\mathcal{B}_{\text{pass}}} \delta\left(y_{iT_l} - \sum_{j\in C_l} y_{ij}\right) \quad (12)$$

where $\mathcal{B}_{\text{pass}}$ denotes the (observed) set of passed balanced edits. Following convention in the optimization-based editing literature (see Riera-Ledesma and Salazar-González 2007, and references therein), the space $\mathcal{Y}\,(\supset \mathcal{X})$ is the predetermined orthotope such that $\mathcal{Y} = \{(y_1, \ldots, y_p) : \tilde{L}_j \le y_j \le \tilde{U}_j, j = 1, \ldots, p\}$, where $(\tilde{L}_j, \tilde{U}_j)$ are constants. In practice these can be set as the minimum and maximum reported values of item j in the industry or the entire census. The range of the observed values $(\tilde{L}_j, \tilde{U}_j)$ may not be the same as the range restriction (L_j, U_j).

When $A_i = 2$, we have

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{s}_i, A_i, \psi_y) = \kappa_i \prod_{\{j:s_{ij}=1, j\in NT\}} \text{Unif}\left(\log y_{ij} \in [\log \tilde{L}_j, \log \tilde{U}_j]\right) \times \prod_{\{j:s_{ij}=0\}} \delta\left(y_{ij} - x_{ij}\right) \prod_{l=1}^{q} \delta\left(y_{iT_l} - \sum_{j\in C_l} y_{ij}\right) I\left[\boldsymbol{y}_i \notin \mathcal{X}\right]. \quad (13)$$

In this case, we require a normalizing constant κ_i that is not straightforward to compute due to the complexity of the feasible region. We approximate it by Monte-Carlo simulation as follows. Step (a): generate $\log y'_{ij}$ from $\text{Unif}\left(\log \tilde{L}_j, \log \tilde{U}_j\right)$ for variables $j \in NT$ such that $s_{ij} = 1$. Step (b): put $y'_{ij} = y_{ij}$ for variables j such that $s_{ij} = 0$. Step (c): fill in the total variables T_l such that $s_{iT_l} = 1$ by $y'_{iT_l} = \sum_{j\in B_l} y'_{ij}$. Step (d): repeat steps (a)–(c) n_{simul} times, where each iterate generates $\boldsymbol{y}_i'^{(r)}$. Step (e): count the number of edit-failing values, $n_{\text{fail}} = \sum_{r=1}^{n_{\text{simul}}} I[\boldsymbol{y}_i'^{(r)} \notin \mathcal{X}]$. Finally, let the normalizing constant for record i be $\kappa_i = n_{\text{simul}}/n_{\text{fail}}$. Each κ_i can be approximated for each $\boldsymbol{s}_i$ prior to MCMC implementation, as it is free from other unobservables including $\boldsymbol{x}_i^F$ and $\boldsymbol{\theta}$. We use $n_{\text{simul}} = 100$ simulated

values in the simulations and CM editing. In the examples, using $n_{\text{simul}} = 1000$ does not result in noticeable improvement in editing and inference.

3.3 Estimating the Model via MCMC

The posterior distribution of interest is $f(\boldsymbol{X}_n, \boldsymbol{S}_n, \Theta \mid \boldsymbol{Y}_n, \boldsymbol{A}_n, \mathcal{X})$, where $\boldsymbol{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{S}_n = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$, $\boldsymbol{Y}_n = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, $\boldsymbol{A}_n = (A_1, \ldots, A_n)$ and Θ includes all parameters in (3) – (13). In this section, we highlight two key aspects of the strategy for sampling $(\boldsymbol{X}_n, \boldsymbol{S}_n, \Theta)$ from the posterior distribution via MCMC. Details of the full MCMC implementation are presented in Appendix A of the supplementary material.

To update $\boldsymbol{\theta}$, we use a data augmentation technique that follows the approach in Kim et al. (2014), which is based on the ideas in O'Malley and Zaslavsky (2008) and Manrique-Vallier and Reiter (2014). We suppose the observed data are a sample from a hypothetical sample of N_{aug} individuals, where N_{aug} is a random variable. We assume that record i in the hypothetical sample $\boldsymbol{X}_{\text{aug}} = (\boldsymbol{X}_n, \boldsymbol{X}_{N_{\text{aug}}-n})$, where $\boldsymbol{X}_{N_{\text{aug}}-n} = \{\boldsymbol{x}_i \notin \mathcal{D} : i = n+1, \ldots, N_{\text{aug}}\}$ are the $N_{\text{aug}} - n$ hypothetical, unobserved individuals, follows a mixture of unconstrained normal distributions given by

$$f(\boldsymbol{x}_i|\mu, \Sigma, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \mathrm{N}\ (\log \boldsymbol{x}_{i,NT}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \prod_{l=1}^{q} \delta\left(x_{iT_l} - \sum_{j \in B_l} x_{ij}\right) \tag{14}$$

for $i = 1, \ldots, N_{\text{aug}}$. Following Meng and Zaslavsky (2002) and O'Malley and Zaslavsky (2008), we assume $n \sim \text{Binomial}\,(N_{\text{aug}}, h(\boldsymbol{\theta}))$, where

$$h(\boldsymbol{\theta}) = \int_{\boldsymbol{x} \in \mathcal{X}} \sum_{k=1}^{K} \pi_k \mathrm{N}(\log \boldsymbol{x}_{NT}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\ d \log \boldsymbol{x}$$

and $f(N_{\text{aug}}) \propto 1/N_{\text{aug}}$. The resulting draws of the $\boldsymbol{\theta}$ from (14) are equivalent to draws from the constrained model in (1).

For each faulty record i, we draw $\boldsymbol{x}_i$ and $\boldsymbol{s}_i$ jointly as follows: (i) propose $\boldsymbol{s}_i'$ consistent with A_i, (ii) propose $\boldsymbol{x}_i'$ given $\boldsymbol{s}_i'$, and (iii) accept or reject the proposed $(\boldsymbol{x}_i', \boldsymbol{s}_i')$ via a Metropolis-Hastings step. Here, we consider only proposed $\boldsymbol{s}_i'$ that can result in feasible solutions under $\mathcal{X}$ given $\boldsymbol{y}_i$. For example, we cannot propose $s_{ij} + s_{ij'} = 0$ when the ratio edit for j and j' fails for that record. When p is small, it is possible to identify each

record's set of feasible $\boldsymbol{s}_i$, which we write as $\mathcal{S}(\boldsymbol{y}_i, A_i) \subset \{0,1\}^p$, before implementing the MCMC sampler. When p is not small, so that enumeration of $\mathcal{S}(\boldsymbol{y}_i, A_i)$ is computationally very expensive, one can propose $\boldsymbol{s}_i'$ and check if it is feasible within the MCMC sampler, which we do via a simplex algorithm; see Appendix B of the supplementary material for details. When p is not small, many feasible solutions are likely to have low probability mass, so that it can be inefficient to propose random jumps within $\mathcal{S}(\boldsymbol{y}_i, A_i)$. Instead, we implement a birth-death process that proposes moves from the current $\boldsymbol{s}_i^{(t-1)}$ to a neighbor in $\mathcal{S}(\boldsymbol{y}_i, A_i)$. The birth-death process results in higher acceptance rates than a completely random proposal which more often proposes moves in regions with low probability mass.

To illustrate the process of sampling $(\boldsymbol{x}_i, \boldsymbol{s}_i)$, we return to the example in Section 2 with two explicit edits, $x_1 \leq x_2$ and $x_2 \leq x_3$, and $\boldsymbol{y}_i = (6, 4, 2)$. Suppose the current value of $\boldsymbol{s}_i^{(t)} = (1, 1, 0)$ in the Markov chain. We propose the next iteration of $\boldsymbol{s}_i$ by changing two values, for example $\boldsymbol{s}_i' = (1, 0, 1)$. We then check whether the proposal is feasible or not using a simplex algorithm, tossing out infeasible combinations such as $\boldsymbol{s}_i = (0, 1, 0)$. Using the proposed $\boldsymbol{s}_i'$, we randomly draw values of $\boldsymbol{x}_i$ for all variables j with $s_{ij}' = 1$, resulting in the proposal $(\boldsymbol{s}_i', \boldsymbol{x}_i')$, where $\boldsymbol{x}_i' = (\boldsymbol{x}_i'^F, \boldsymbol{y}^{UF})$. The Metropolis-Hastings step accepts or rejects $(\boldsymbol{s}_i', \boldsymbol{x}_i')$ by comparing the conditional densities $(\boldsymbol{s}_i, \boldsymbol{x}_i | \mu, \Sigma, z_i, \boldsymbol{y}_i, \psi_s, \psi_y, A_i)$ of $(\boldsymbol{s}_i^{(t)}, \boldsymbol{x}_i^{(t)})$ and $(\boldsymbol{s}_i', \boldsymbol{x}_i')$. We do not need to derive the implied edits to check the feasibility of $\boldsymbol{s}_i'$ given $\boldsymbol{y}_i$. We can use the simplex algorithm in standard form to check whether the suggested $\boldsymbol{s}_i$ generates a non-null space of $\boldsymbol{x}_i$ given the $\boldsymbol{y}_i$ and explicit edits only.

Generally, the MCMC sampler does not visit all or even most feasible solutions of $(\boldsymbol{x}_i, \boldsymbol{s}_i)$. Rather, the Metropolis Hastings encourages the model to find and explore feasible regions of support with high density under the assumed model. Enumerating all possible solutions is not required for this exploration.

4. SIMULATION STUDIES

In this section, we present results of a simulation study comparing the Bayesian editing model to F-H and other approaches. We generate a population $\boldsymbol{X}^{\text{pop}}$ comprising 1,000,000 records measured on $p = 9$ variables. Each $\boldsymbol{x}_i$ satisfies two balance edits, $x_{i1} = x_{i2} + x_{i3} + x_{i4}$ and $x_{i5} = x_{i6} + x_{i7}$, and ratio edits for all pairs of variables of interest in $\{x_{i1}, x_{i5}, x_{i8}, x_{i9}\}$.

This mimics the structure of the constraints in the CM, namely some variables of interest must satisfy both balance and ratio edits and other variables of interest must satisfy only ratio edits. We generate each $\boldsymbol{x}_i$ by sampling all variables except x_{i1} and x_{i5} from a mixture of three multivariate normal distributions and then setting x_{i1} and x_{i5} equal to sums of their component variables. Each $\boldsymbol{x}_i \in \boldsymbol{X}^{\text{pop}}$ satisfies all edit constraints; see Appendix C of the supplementary material for details of edit rules and parameter values.

We sample $R = 500$ independent simple random samples $\boldsymbol{X}^r$, where $r = 1, \ldots, R$, of size $n = 1000$ from $\boldsymbol{X}^{\text{pop}}$. In each $\boldsymbol{X}^r$, we randomly select 600 records to have $A_i = 0$ for which $\boldsymbol{x}_i = \boldsymbol{y}_i$, 200 records to have $A_i \in \{1, 3\}$, and 200 records to have $A_i = 2$. Within each group of 200 records, we randomly draw the simulated values of $\boldsymbol{s}_i$ and $\boldsymbol{y}_i$ by rejection sampling: generate $s_{ij} \sim \text{Bernoulli}(\psi_{s_j})$ for $j = 1, \ldots, p$, generate y_{ij} for $s_{ij} = 1$ from the appropriate uniform distribution, and accept the generated values when $\boldsymbol{y}_i$ violates edit rules in accord with the corresponding A_i. See Appendix C of the supplementary material for parameter values ψ_{s_j}.

In each simulation run, we implement several methods for data editing, including

- `BE`: Bayesian editing described in Section 3.
- `FH`: F-H error localization under a MWFI criterion, and imputation of flagged variables from the model in Section 3 fixing each $\boldsymbol{s}_i$ at the F-H solution. We set the reliability weight $w_j = 1 - \sum_{i=1}^{n} s_{ij}^*/n$, where $\boldsymbol{s}_i^*$ is the simulated (true) $\boldsymbol{s}_i$ corresponding to $\boldsymbol{y}_i$.
- `BE-min`: Bayesian editing under a MFI criterion, i.e, restricting the support of $\boldsymbol{s}$ to the set of values that minimize $\sum_j s_{ij}$.
- `AAI`: Fixing $s_{ij} = 1$ for all active items, i.e., all variables involved in edit violations, and imputation of flagged variables from the model in Section 3.
- `BE-sgl`: Bayesian editing with a single, constrained multivariate normal distribution for $f(\boldsymbol{x}_{i,NT})$ instead of the mixture of K multivariate normal distributions.

For each method, we run an appropriate MCMC algorithm with 5000 iterations after a burn-in period of 5000 iterations. We keep the imputed values of each $\boldsymbol{x}_i^F$ for every 500th

iteration to create $M = 10$ multiply imputed, plausible datasets $\{\boldsymbol{X}^{r(1)}, \ldots, \boldsymbol{X}^{r(M)}\}$. See Appendix D of the supplementary material for the MCMC steps for editing methods other than `BE`.

In general, across simulation runs the completed datasets from `BE` tend to approximate the distribution of $\boldsymbol{X}^r$ more faithfully than the other editing methods. This is illustrated in Figure 2, which displays plots of $\log x_{i1}$ and $\log x_{i9}$ for one $\boldsymbol{X}^r$ and one randomly selected completed dataset $\boldsymbol{X}^{r(m)}$. Unlike the distribution of $\boldsymbol{X}^{r(m)}$ for `BE`, the distributions for `FH` and `BE-min` include imputed values that have relatively low probability density according to the distribution of $\boldsymbol{X}^r$ on the log scale. This is similar to the pattern seen in Figure 1. We note that `AAI` results in reasonable distributions, with added variability in imputations compared to `BE` due to the replacement of additional values. Similar patterns are evident in Figure 3, which displays the pairwise correlations across all variables in $\boldsymbol{X}^r$ and each $\boldsymbol{X}^{r(m)}$. The correlations for `BE` are quite similar to those in $\boldsymbol{X}^r$, whereas `FH` and `BE-min` result in underestimation.

We also investigate repeated sampling properties of the editing procedures. After creating the $M = 10$ datasets for each simulation, we apply standard multiple imputation techniques to calculate point estimates and 95% confidence intervals for a variety of population quantities. We also compute these quantities for three models that offer context on the quality of the editing methods. These include using $\boldsymbol{X}^r$ for all cases, using the sample of edit-passing records only ($\boldsymbol{X}^{r,\text{pass}}$) as a type of "complete-case" analysis, and using the model in Section 3 with each s_{ij} fixed at the true s^*_{ij}. The first and third analyses generally are not possible outside simulation studies.

Let Q denote some population quantity in $\boldsymbol{X}^{\text{pop}}$, and let $\hat{q}_r$ be the point estimate of Q from replicate sample r. For each method, we compute relBias $= \left(\sum_{r=1}^{R} \hat{q}_r/R - Q\right)/|Q|$, and relRMSE $= \left(\sqrt{\sum_{r=1}^{R}(\hat{q}_r - Q)^2/R}\right)/|Q|$. We also compute the empirical percentage of 95% confidence intervals that include their corresponding Q. As representative Q, we use the mean of each variable and the coefficients in the regression,

$$\log x_{i9} = \beta_0 + \beta_1 \log x_{i1} + \beta_2 \log x_{i5} + \beta_3 \log x_{i8} + \varepsilon_i, \;\; \varepsilon_i \sim \text{N}(0, \sigma^2).$$

Population values for the means and βs are computed from $\boldsymbol{X}^{\text{pop}}$.

Tables 1 and 2 summarize the simulation results for the population means and regres-

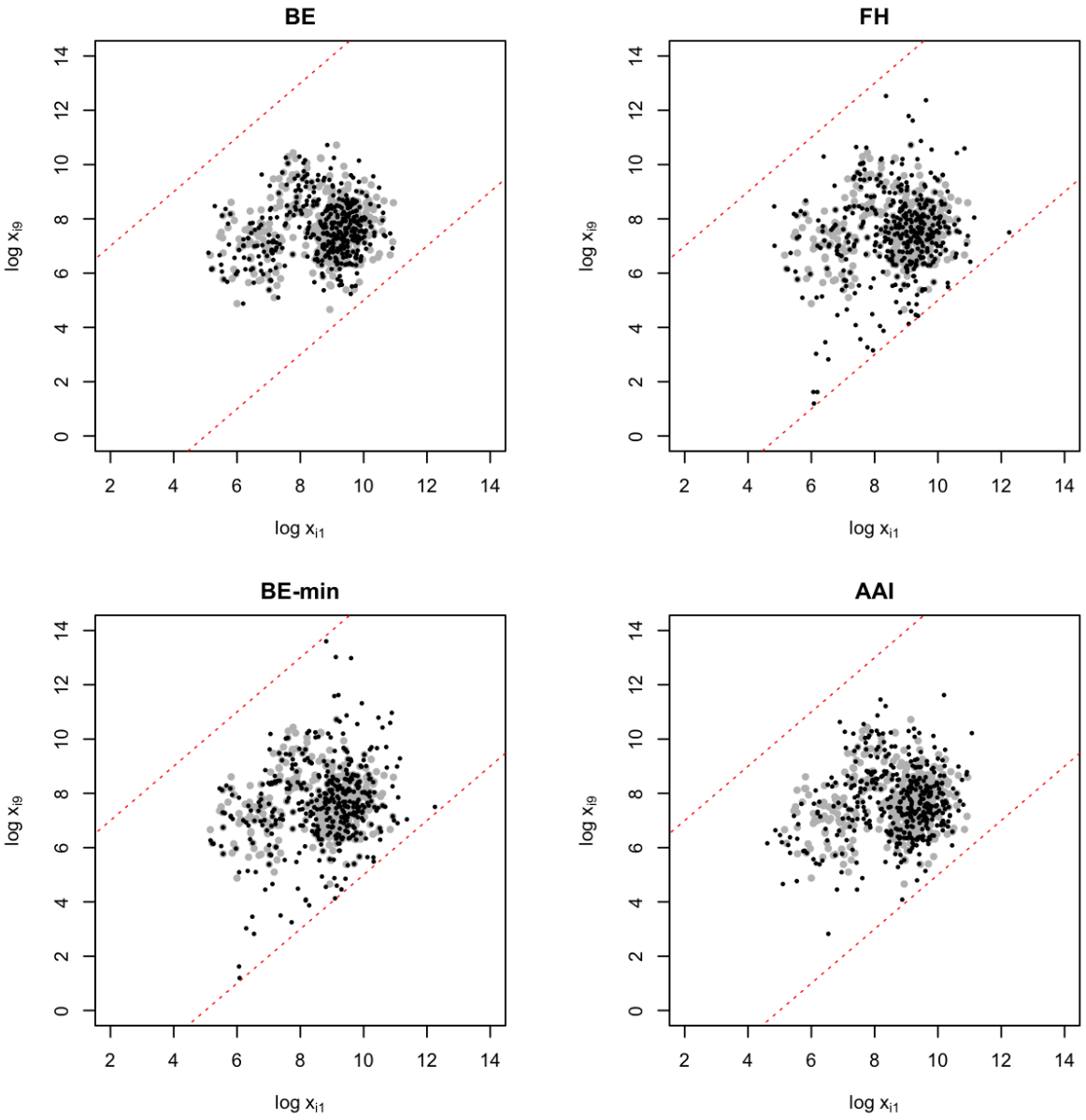

Figure 2: Plots of $\log x_{i1}$ and $\log x_{i9}$ in the simulation example. The black dots represent $\boldsymbol{X}^{r(m)}$ of `BE` (top left), `FH` (top right), `BE-min` (bottom left) and `AAI` (bottom right); the gray dots in background represent $\boldsymbol{X}^{r}$.

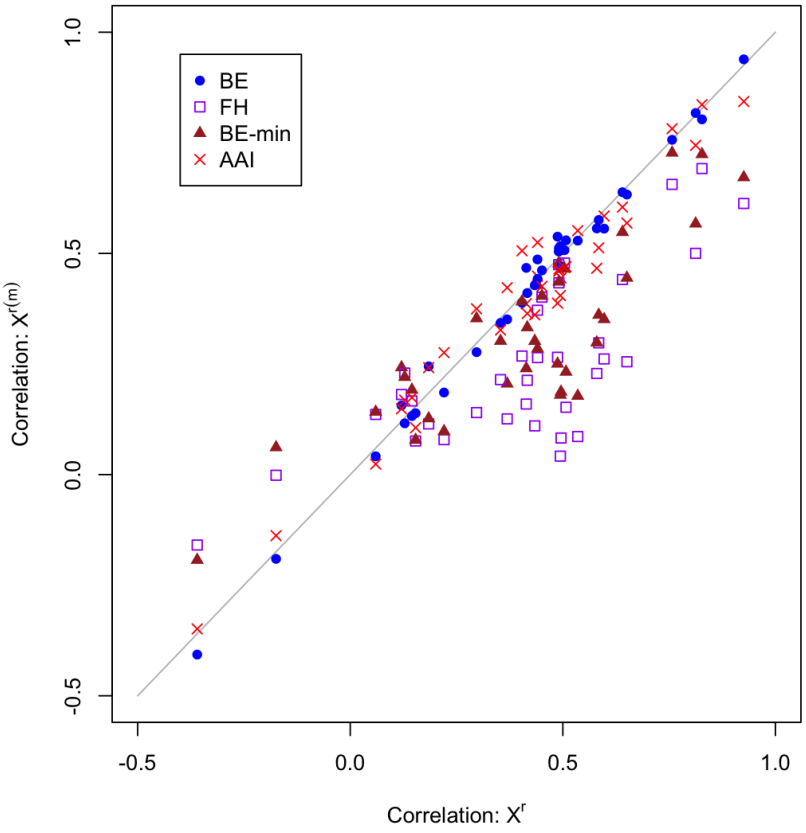

Figure 3: The set of pairwise correlations from the nine log transformed variables. X-axis and Y-axis represent the correlation coefficients calculated from a true sample $\boldsymbol{X}^r$ and an edited dataset $\boldsymbol{X}^{r(m)}$, respectively. The solid dots of `BE` which are close to the 45 degree line indicate that `BE` preserves the correlation structure of $\boldsymbol{X}^r$.

Table 1: Summaries of the estimators of population mean across $R = 500$ simulations. The columns labeled $\boldsymbol{X}^r$ and $\boldsymbol{X}^{r,\text{pass}}$ display results based on the true data and the edit-passing records only, respectively. The column labeled `TrueS` displays the results using the model in Section 3 with each s_{ij} fixed at the true s^*_{ij}.

				Editing methods				
	$\boldsymbol{X}^r$	$\boldsymbol{X}^{r,\text{pass}}$	TrueS	BE	FH	BE-min	AAI	BE-sgl
relBias (×100)								
$\bar{X}_1$	0.1	0.3	0.1	-0.1	2.8	6.2	0.9	3.4
$\bar{X}_2$	0.6	0.1	2.1	0.4	65.0	88.3	8.3	-5.2
$\bar{X}_3$	0.2	0.4	0.2	0.1	1.5	3.3	0.2	3.5
$\bar{X}_4$	0.0	0.0	-0.1	-0.6	0.4	5.5	1.7	4.0
$\bar{X}_5$	-0.1	-0.1	-0.1	-1.1	103.9	91.3	6.5	-0.7
$\bar{X}_6$	-0.1	-0.2	0.0	-0.9	213.0	168.6	8.8	0.0
$\bar{X}_7$	-0.1	0.0	-0.2	-1.2	56.0	57.4	5.4	-1.0
$\bar{X}_8$	0.4	0.5	0.5	0.7	2.3	6.6	0.7	2.0
$\bar{X}_9$	0.0	-0.2	0.4	-0.5	18.6	31.2	4.3	-1.5
relRMSE (×100)								
$\bar{X}_1$	2.7	3.4	2.7	2.8	4.5	7.6	3.3	4.5
$\bar{X}_2$	8.1	10.0	8.6	8.4	68.4	94.8	14.2	10.1
$\bar{X}_3$	3.5	4.5	3.5	3.6	4.9	6.7	4.3	5.2
$\bar{X}_4$	3.7	4.8	3.8	4.0	4.1	7.7	4.9	6.0
$\bar{X}_5$	2.4	3.1	2.4	2.7	107.0	96.2	8.1	2.8
$\bar{X}_6$	3.3	4.2	3.4	3.7	222.6	185.7	13.3	4.5
$\bar{X}_7$	2.8	3.7	2.9	3.2	60.4	65.6	7.5	3.3
$\bar{X}_8$	2.8	3.6	3.0	3.1	4.4	8.2	3.4	3.8
$\bar{X}_9$	4.5	5.9	4.8	4.8	21.5	36.3	7.8	5.3
95% CI Coverage								
$\bar{X}_1$	95.2	95.4	96.2	95.8	90.0	73.8	96.2	89.2
$\bar{X}_2$	93.0	95.4	95.6	95.4	6.4	15.4	97.0	86.6
$\bar{X}_3$	94.4	95.6	94.0	96.2	95.2	96.4	97.6	92.6
$\bar{X}_4$	93.4	93.0	94.6	94.8	96.6	87.0	95.2	93.0
$\bar{X}_5$	93.8	94.0	94.4	92.4	0.0	2.0	93.4	94.2
$\bar{X}_6$	94.8	94.2	93.8	93.0	0.8	23.4	97.8	94.8
$\bar{X}_7$	94.8	94.4	94.2	92.2	10.8	42.2	94.4	93.0
$\bar{X}_8$	95.0	95.6	94.6	93.8	96.6	84.8	95.8	93.0
$\bar{X}_9$	95.6	92.2	96.4	95.4	67.0	56.8	94.0	92.4

sion coefficients. Overall results across the two tables are similar. Among all the editing methods, `BE` tends to result in the lowest biases and root mean squared errors (RMSE), while having close to nominal 95% coverage rates. In contrast, `FH` and `BE-min` are highly unreliable, resulting in some estimates with high bias and poor coverage rates. `AAI` also tends to result in reasonable coverage rates, although it tends to have larger RMSEs than `BE` (but smaller RMSEs than `FH` and `BE-min`). The RMSEs for `BE` tend to be smaller than those based on $\boldsymbol{X}^{r,\text{pass}}$, indicating that `BE` takes advantage of information in the faulty cases that complete-case analysis would ignore. All editing methods have diminished performance compared to using $\boldsymbol{X}^{r}$, which is to be expected.

The results also suggest that the choice of error localization technique is crucial to the performance of editing procedures. The RMSEs for `TrueS` are close to those based on $\boldsymbol{X}^{r}$, indicating that the imputation model underlying the editing strategies is effective—this is not too surprising since the imputation model matches the data generation model. The RMSEs of `BE` are only slightly worse than those for `TrueS`, reflecting the merits of stochastic error localization informed by the data relative to MFI (MWFI) error localization procedures or an all active items procedure. Interestingly, using an incorrect imputation model with stochastic error localization (`BE-sgl`) still outperforms the procedures based on MFI (MWFI) error localization even with a correct imputation model.

Over the 500 replications, the averages of the number of changed fields for cases with edit failures, $\sum_{\{i:A_i>0\}}\sum_j s_{ij}/400$, are 2.8 for `TrueS`, 1.7 for `FH`, and 7.0 for `AAI`. For methods updating of $\boldsymbol{s}_i$ during MCMC iterations, the averages of $\sum_{\{i:A_i>0\}}\sum_j s_{ij}/400$ over the 10 multiply imputed datasets and 500 replications are 3.8 for `BE`, 1.7 for `BE-min`, and 4.0 for `BE-sgl`. Thus, `FH` and `BE-min` tend to underestimate the number of variables that should be replaced, whereas `BE` (and `BE-sgl`) tend to overestimate the number. Evidently, with a good-fitting imputation model, changing too many variables is less problematic than changing too few variables, the latter of which can result in unrealistic imputed values.

We also evaluated data quality using the propensity score metric from the literature on statistical disclosure limitation (Woo et al. 2009), which determines how well one can discriminate $\boldsymbol{X}^{r(m)}$ from $\boldsymbol{X}^{r}$. Inability to discriminate suggests similar distributions, which means high quality of $\boldsymbol{X}^{r(m)}$. Results were similar to those in Tables 1 and 2; these are

Table 2: Summaries of the estimators of regression coefficients across $R = 500$ simulations. The columns labeled $\boldsymbol{X}^r$ and $\boldsymbol{X}^{r,\text{pass}}$ display results based on the true data and the edit-passing records only, respectively. The column labeled `TrueS` displays the results using the model in Section 3 with each s_{ij} fixed at the true s^*_{ij}.

				Editing methods				
	$\boldsymbol{X}^r$	$\boldsymbol{X}^{r,\text{pass}}$	`TrueS`	`BE`	`FH`	`BE-min`	`AAI`	`BE-sgl`
relBias (×100)								
β_0	0.2	0.1	0.3	0.9	-2.6	2.8	-1.8	2.5
β_1	-0.8	-1.6	-0.3	-2.9	51.7	39.7	10.3	-5.4
β_2	0.0	0.4	0.3	1.7	-41.6	-24.7	-3.3	3.1
β_3	0.2	0.5	-0.3	-0.4	0.0	-9.0	-2.2	-1.4
relRMSE (×100)								
β_0	8.0	10.2	8.9	9.4	12.9	12.3	10.5	9.9
β_1	13.7	18.2	15.0	16.3	54.4	43.2	21.2	17.0
β_2	9.9	12.9	10.9	11.9	43.2	28.0	13.3	12.3
β_3	9.2	12.0	10.4	11.0	13.5	15.6	11.7	11.3
95% CI Coverage								
β_0	94.8	94.6	94.0	95.2	88.6	89.4	94.0	93.2
β_1	93.2	91.8	92.8	94.0	11.8	34.6	91.4	92.4
β_2	92.6	94.0	92.8	93.2	3.8	37.8	95.2	92.6
β_3	94.4	93.2	95.2	94.2	91.0	83.0	93.2	92.4

reported in Appendix E of the supplementary material.

We also ran simulation studies in which $f(\boldsymbol{s}_i, A_i \mid \boldsymbol{x}_i, \psi_s) \neq f(\boldsymbol{s}_i, A_i \mid \psi_s)$; results are reported in Appendix F of the supplementary material. In particular, we ran a FAR scenario in which the data generating distribution for A_i depends on one variable known to be correct for all records, and a not FAR scenario in which the the data generating distribution for A_i depends on a variable that is subject to reporting errors. We fit the editing procedures without any adjustments. In the additional FAR scenario, the results for the editing procedures are similar to those in Tables 1 and 2, with `BE` the only editing procedure that offers reliable inferences. In the not FAR scenario, all editing procedures result in some obviously invalid inference, as expected since each model presumes some variant of FAR. Interestingly, in this simulation `BE` continues to dominate the other editing procedures.

Finally, since some statistical agencies use outlier detection algorithms before implementing F-H error localization, we also run a simulation comparing `BE` to such an approach; see Appendix G in the supplementary materials for details. In particular, we identify univariate outliers using a technique from the Banff system of Statistics Canada (Kozak 2005; Banff Support Team 2007), examining three cutpoints for defining outliers. We then set $s_{ij} = 1$ for cases identified as outliers, and run a minimum number of fields to impute approach (`BE-min`, since it outperformed `FH`) on the subsequent data. For some cutpoint values, forced editing of the selected outliers improves the quality of inferences for `BE-min`. However, `BE` still dominates `BE-min` combined with outlier detection at each cutpoint. We also examined a procedure that sets $s_{ij} = 1$ for cases identified as outliers before running the `BE` approach. This modified version of `BE` also outperforms `BE-min` with outlier detection, suggesting benefits of stochastic error localization even after outlier identification.

5. EDITING THE CENSUS OF MANUFACTURES

The Census of Manufactures is part of the U.S. Economic Census, which is conducted every five years by the Census Bureau. The CM data comprise information on manufacturing establishments in the U.S. with one or more paid employees, including data on employment, payroll, and production. To collect CM data, the Census Bureau mails a questionnaire

to every manufacturing establishment except so-called administrative record cases, which generally are very small manufacturing establishments (typically fewer than five employees). Values for these establishments are taken from administrative records of other federal agencies. The CM data are subject to Title 13 and hence accessible only to approved researchers in the Census Research Data Centers. We use data from the 2007 Economic Census, as data from the 2012 Economic Census were not available to us. We note that the results below are primarily illustrative of automatic editing, as there are additional complexities in the CM that are not addressed here.

The CM data are organized by industry classification codes. Within each code, establishments are required to satisfy industry-specific edit rules (Winkler and Draper 1996; Garcia and Goodwin 2002; Thompson et al. 2004). Inevitably, the reported data violate these rules, so that the Census Bureau must do edit-imputation. For some establishments, particularly those known to be large, the Census Bureau uses telephone interviews to collect information on missing values and to verify or correct values suspected to be faulty. It also takes some information from administrative records. Due to cost considerations, the Census Bureau does not use these manual edit procedures for all establishments. For the remaining establishments, the Census Bureau uses the SPEER system (Greenberg and Surdi 1984; Winkler and Draper 1996; Draper and Winkler 1997) for automatic editing. This system implements a variant of MWFI error localization followed by a combination of mean imputation, ratio imputation, and regression imputation.

The CM data include $p = 27$ variables involved in edits, comprising 12 variables subject to ratio edits, of which six also are subject to balance edits, and 15 component variables that are part of balance edits but not subject to ratio edits. All 27 variables and the balance edits are described in Table 3; see Appendix H of the supplementary material for more information on the edit rules. We replace the balance edit $\text{TE} = \text{PW} + \text{OE}$ with $\text{TE} = (\text{PW}_1 + \text{PW}_2 + \text{PW}_3 + \text{PW}_4)/4 + \text{OE}$, that is, we combine the nested balance edits for total variables TE and PW.

We use an industry broadly described as metalworking machinery manufacturing (NAICS code of 33351400). Data for this industry were made available to us by the Census Bureau. The CM data for this industry include 1869 establishments. These data already have been

Table 3: Variables used in the CM edit rules. All 12 variables in the first column are subject to ratio edits. Here, C_*, IB_* and IE_* denote reported components for TC, TIB and TIE, respectively, and PW_* denotes the number of production workers reported in each quarter.

Variables	Description	Components (if any)
TC	Total Cost of Materials ($1000)	$C_a + C_b + C_c + C_d + C_e$
TIB	Total Inventory Begin Yr. ($1000)	$IB_a + IB_b + IB_c$
TIE	Total Inventory End Yr. ($1000)	$IE_a + IE_b + IE_c$
TVS	Total Value of Shipments ($1000)	
PW	Number of Production Workers	$(PW_1 + PW_2 + PW_3 + PW_4)/4$
OE	Number of Other Employees	
TE	Total Employment	PW + OE
WW	Production Workers Wages ($1000)	
OW	Other Workers Wages ($1000)	
SW	Total Salaries and Wages ($1000)	WW + OW
BEN	Total Benefits to Employees ($1000)	
PH	Production Worker Hours (1000 hours)	

subject to manual edits and are the same as those edited using the SPEER system. We replace zeros in $\boldsymbol{y}_i$ with $\omega = 0.1$. This is necessary for ratio edits to be sensible, since one cannot divide by zero (see Kovar et al. 1988). Out of the 1869 records, 585 records fail the edits, either due to missing values or violations in the reported values. Specifically, 489 records fail at least one balance edit ($A_i \in \{1, 3\}$) and 96 records pass all balance edits but fail ratio edits ($A_i = 2$).

We apply `BE` and `BE-min` to create two sets of $M = 10$ corrected versions of $\boldsymbol{X}$ that satisfy all edit constraints. For each, we run a single MCMC for 10000 iterations, tossing the first 5000 as burn in. We then select M datasets by storing every 500th iterate. The 10000 iterations take approximately 18 hours of CPU time using a server available in the Research Data Centers of the Census Bureau. To check convergence of the MCMC chain, we monitor the draws of N_{aug} and α. These parameters are not subject to label switching and, in our experience, are highly sensitive to lack of convergence.

The resulting differences in the completed datasets are illustrated in Figure 4, which dis-

plays multiple imputation estimates of pairwise correlations among the 27 log-transformed variables after edit-imputation versus the corresponding correlations based on only the edit-passing records ($\boldsymbol{X}^{r,\text{pass}}$). The correlations from `BE` are similar to those computed from $\boldsymbol{X}^{r,\text{pass}}$, whereas the correlations from `BE-min` are noticeably attenuated. This follows the pattern in Figure 3 from the simulation study. We note that the correlations from the SPEER edits are similar to those from `BE-min` (although a handful of correlations are completely different because the Census Bureau subsequently replaced values of some variables with imputations using information not available to us). Assuming the mechanism for the faulty/incomplete values is ignorable, one would expect imputations for faulty data to reflect the correlational structure in the edit-passing data. Although we cannot be certain that `BE` results in more plausible imputations than the MFI approach—as with any missing data setting, where the truth is not known—the closer similarity with the complete case results is suggestive that `BE` offers more plausible imputations. We note that using the `BE` imputations carries advantages over using the edit-passing records only, in that (i) deleting partially completed cases sacrifices efficiency compared to editing and imputing the faulty cases, and (ii) using only the complete cases creates terrible biases when estimating (unweighted) population totals.

`BE` and `BE-min` also result in completed datasets with different means, as evident in Figure 5. The means from `BE` generally are smaller than those from `BE-min`. Once again, this pattern accords with the simulation results in Table 1.

6. CONCLUDING REMARKS

Our simulation studies suggest that using stochastic error localization can result in better performance than using the minimum number of fields criterion. Stochastic error localization allows relationships in the data to inform the localization, while fully reflecting uncertainty. The model-based approach always generates corrected values that satisfy all edits; only a single pass through the data is required. The approach also avoids the computationally difficult exercise of identifying all implied edits used in some automatic editing processes (see Sande 1979; Garcia 2002). The support of the imputation step is pre-defined only on the feasible space, and proposed solutions are checked via a simplex method embed-

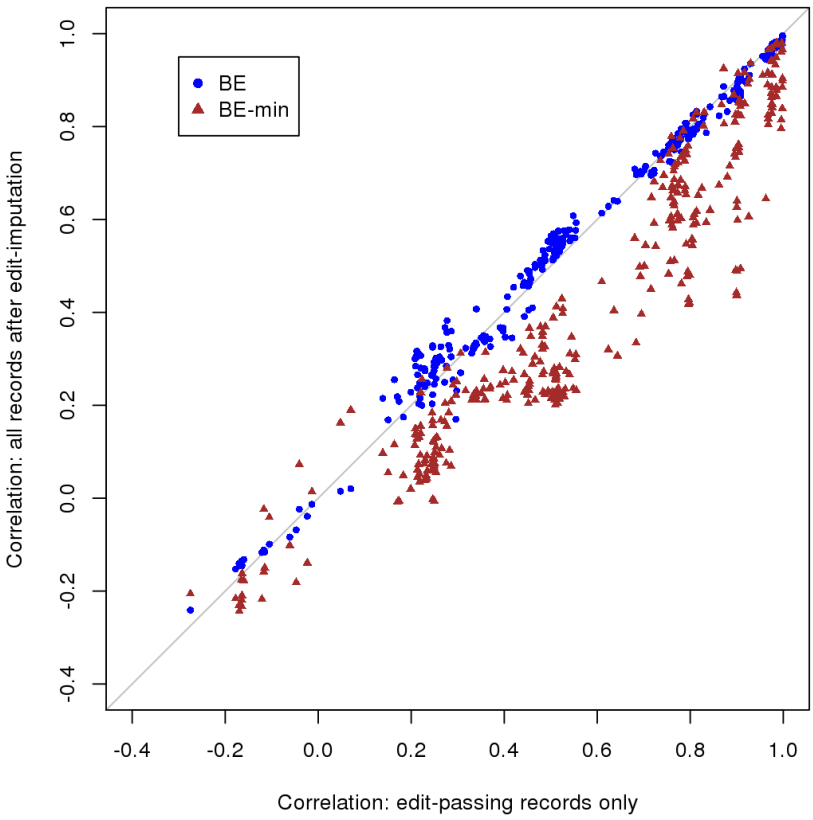

Figure 4: Pairwise correlations from the 27 log-transformed variables in the CM data computed for edit-passing records only (X-axis) and for all records after edit-imputation (Y-axis).

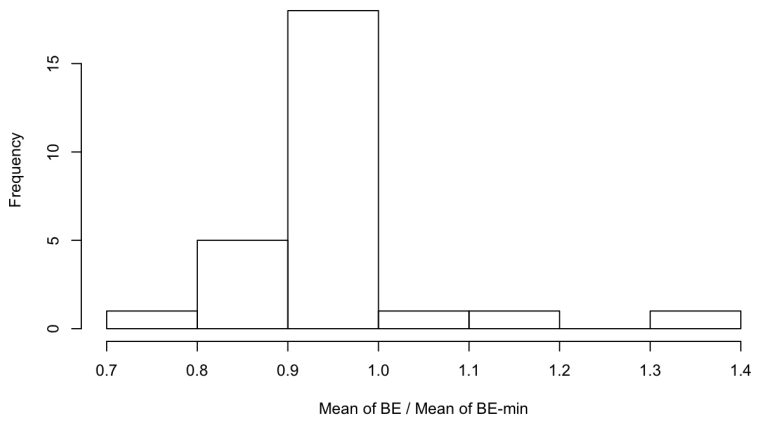

Figure 5: Ratios of means of `BE` to those of `BE-min` for the 27 variables of the CM data. Values below 1.0 imply that the mean of `BE` is smaller than the mean of `BE-min`.

ded in the estimation algorithms. The approach handles both balance edits and inequality constraints simultaneously.

As noted by reviewers, agencies using F-H approaches could construct heuristics that force higher probability of editing certain fields based on combinations of variables from other fields. For example, using the motivating example of a 40 year old pregnant male in Section 1, the agency could set a rule that changes pregnancy status when the 40 year old person also subscribes to men's magazines and buys men's clothing, but change gender when the 40 year old buys diapers and maternal clothing. Such heuristics could get cumbersome in high dimensions, particularly when heuristics should be based on multivariate relationships. The stochastic editing approach automatically lets the data identify unusual combinations based on relationships among all variables in the data, thereby potentially leveraging important patterns that were not pre-determined by the agency. Unlike F-H approaches with such heuristics, the stochastic editing approach allows for uncertainty in the fields to be imputed. The 40 year old person still could be a pregnant woman buying men's clothes for her husband (or herself) and subscribing to men's magazines, or a man buying diapers for his child and clothes for his wife. We note that, as suggested by one reviewer, agencies might introduce uncertainty in F-H approaches by creating multiple edited datasets from near-optimal solutions; we are unaware of research or implementations

of this potential approach.

We believe that the Bayesian editing approach can be applied "as is" in many settings with edit constraints based on equalities and linear inequalities. Mixtures of multivariate normal distributions are flexible enough to capture many distributional shapes with minimal tuning, and the uniform distributions for the reporting errors and the error indicators represent a default position for lack of knowledge about the nature of the measurement error. An R package implementing Bayesian editing for continuous data will be available on CRAN in early 2015. That said, agencies could adjust aspects of the model to their advantage; for example, replace the uniform distribution for reporting errors with an informative distribution determined from past experience. For individuals experienced with Bayesian modeling, the adaptations to the Gibbs sampler are straightforward.

While the MCMC algorithm we coded runs in reasonable time on the CM data, stochastic editing can be implemented with substantially increased computational efficiency. Each $(\boldsymbol{x}_i, \boldsymbol{s}_i)$ can be updated in parallel with very little overhead cost to manage the parallel threads. Parameters from the Dirichlet process mixture model can be updated using efficient parallel computation algorithms developed for cluster and GPU computing Suchard et al. (2010). It is also possible to simplify the model to reduce computing time, for example by reducing the number of clusters K or by capping the maximum total number of fields that the model allows to be in error (e.g., require $\sum s_{ij} \leq m$ for some m). This latter strategy reduces the space of plausible $\boldsymbol{s}_i$ to search over. When estimating one model for each of many industry types, the agency can distribute the models over many CPUs. For MCMC convergence diagnostics in such cases, we suggest running long chains and focusing evaluations on industries with large proportions of edit failures, as these are most likely to have the most trouble with convergence.

The favorable performance of the Bayesian editing model suggests several areas for future research. First, due to lack of knowledge about measurement error in the Census of Manufactures, we use uniform distributions for the reporting and error indicator models. Conceptually, it is straightforward to incorporate other reporting models or error indicator models that are functions of missing x_{ij}. It would be informative to characterize how much information can be gained when using informative reporting and error indicator distribu-

tions versus these simple models. Such investigations also could shed light on when it is worthwhile for agencies to spend resources on specifying measurement error models. Second, for values involved in edit failures, the Bayesian editing model encourages corrections that are consistent with the distributions of the edit-passing values. For cases with odd relationships among their true values, the edited values could be shrunk inappropriately towards the mass of edit-passing data. It would be informative to evaluate the impact of such shrinkage on the quality of inferences, both for the Bayesian editing model and for MFI solutions. We conjecture that it will be advantageous to use manual editing for unusual cases. Of course, budgets for manual edits are finite, which points to another topic for research: how organizations should select the records to edit manually also known as selective editing (De Waal 2013; Arbues et al. 2013; Di Zio and Guarnera 2013; Pannekoek et al. 2013). Third, the use of a uniform distribution for the reported values greatly simplifies the treatment of missing y_{ij}, as the model does not need $\boldsymbol{y}_i^F$ once $\boldsymbol{s}_i$ is determined and $s_{ij} = 1$ for all cases with missing reported values (as an aside, we recommend including flags in any released data that differentiate missing and erroneous fields, so as to help secondary analysts understand the corrected data). Clearly, if the missing x_{ij} follow different distributions than the x_{ij} subject to errors, treating missing and erroneous data identically will result in lower quality inferences. It would be useful to develop methods for handling such situations. Finally, for many economic datasets, agencies generate and disseminate tables of aggregates from the edited microdata. Within any completed dataset resulting from the Bayesian editing procedure, all edited values and hence tabulated summaries can be made integers. However, when averaged across the multiple completed datasets, some point estimates of aggregates are likely to be non-integers. Further research is needed on how to ensure integer-valued aggregates after multiple imputation (of any sort), and on whether integers are even necessary in the context of a model that explicitly recognizes the uncertainty due to edit-imputation.

SUPPLEMENTARY MATERIAL

Appendices: The steps for the MCMC algorithm for the Bayesian editing model, methodology for checking whether or not a proposed error localization has a feasible solution,

specification of the parameters used in the simulation study, summaries of fitting procedures for the other editing methods used in the simulation study, an evaluation of the quality of the edited data in the simulation study via a propensity score utility measure, additional simulation studies where A_i depend on the values of $\boldsymbol{x}_i$, additional simulation studies with outlier detection methods, and information on the edits for the industry we use in the CM application. (.pdf file)

Notations: Table of notations and abbreviations used in the main text. (.pdf file)

REFERENCES

Arbues, I., Revilla, P., and Salgado, D. (2013), "An Optimization Approach to Selective Editing," *Journal of Official Statistics*, 29, 489–510.

Banff Support Team (2007), "Functional Description of the Banff System for Edit and Imputation, Version 2.02 ," Technical Report, Statistics Canada.

Barcaroli, G., and Venturi, M. (1997), "DAISY (Design, Analysis and Imputation System): Structure, Methodology and First Applications," in *Statistical Data Editing, Vol. 2: Methods and Techniques*, eds. J. Kovar and L. Granquist, United Nations Publications, pp. 40–50.

Biemer, P. P. (2010), "Total Survey Error: Design, Implementation, and Evaluation," *Public Opinion Quarterly*, 74, 817–848.

De Jonge, E., and Van der Loo, M. (2011), "Manipulation of Linear Edits and Error Localization With the Editrules Package," Discussion Paper 201120, Statistics Netherlands, The Hague/Heerlen.

——— (2014), "Error Localization as a Mixed Integer Problem With the Editrules Package," Discussion Paper 201407, Statistics Netherlands, The Hague/Heerlen.

De Waal, T. (1996), "CherryPi: A Computer Program for Automatic Edit and Imputation," presented at UN/ECE Work Session on Statistical Data Editing, Voorburg.

——— (2013), "Selective Editing: A Quest for Efficiency and Data Quality," *Journal of Official Statistics*, 29, 473–488.

De Waal, T., and Coutinho, W. (2005), "Automatic Editing for Business Surveys: An Assessment of Selected Algorithms," *International Statistical Review*, 73, 73–102.

De Waal, T., and Quere, R. (2003), "A Fast and Simple Algorithm for Automatic Editing of Mixed Data," *Journal of Official Statistics*, 19, 383–402.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*, Hoboken, NJ: John Wiley & Sons.

Di Zio, M., and Guarnera, U. (2013), "A Contamination Model for Selective Editing," *Journal of Official Statistics*, 29, 539–555.

Draper, L. R., and Winkler, W. E. (1997), "Balancing and Ratio Editing With the New SPEER System," Research Report RR97/05, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

Dunson, D. B., and Xing, C. (2009), "Nonparametric Bayes Modeling of Multivariate Categorical Data," *Journal of the American Statistical Association*, 104, 1042–1051.

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

Garcia, M. M. (2002), "Implied Edit Generation and Error Localization for Ratio and Balancing Edits," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 1122–1127.

Garcia, M. M., and Goodwin, R. (2002), "Developing SAS Software for Generating a Complete Set of Ratio Edits," Research Report RRS2002/06, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

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

——— (1988), "Error Localization for Erroneous Data: Continuous Data, Linear Constraints," *SIAM Journal on Scientific and Statistical Computing*, 9, 922–931.

Ghosh-Dastidar, B., and Schafer, J. L. (2003), "Multiple Edit/Multiple Imputation for Multivariate Continuous Data," *Journal of the American Statistical Association*, 98, 807–817.

Granquist, L., and Kovar, J. G. (1997), "Editing of Survey Data: How Much Is Enough?" in *Survey Measurement and Process Quality*, eds. L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dipp, N. Schwarz, and D. Trewin, New York: Wiley, pp. 415–435.

Greenberg, B. G., and Surdi, R. (1984), "A Flexible and Interactive Edit and Imputation System for Ratio Edits," Research Report RR84/18, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

Groves, R. M. (1989), *Survey Errors and Survey Costs*, Hoboken, NJ: John Wiley & Sons.

Groves, R. M., and Lyberg, L. (2010), "Total Survey Error: Past, Present, and Future," *Public Opinion Quarterly*, 74, 849–879.

Ishwaran, H., and James, L. F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161–173.

Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H., and Karr, A. F. (2014), "Multiple Imputation of Missing or Faulty Values Under Linear Constraints," *Journal of Business & Economic Statistics*, 32, 375–386.

Kovar, J., Whitridge, P., and MacMillan, J. (1988), "Generalized Edit and Impuation System for Economic Surveys at Statistics Canada," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 627–630.

Kozak, R. (2005), "The Banff System for Automated Editing and Imputation," in *Proceedings of the Survey Methods Section*, pp. 1–10.

Lavine, M., and West, M. (1992), "A Bayesian Method for Classification and Discrimination," *Canadian Journal of Statistics*, 20, 451–461.

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data* (2nd ed.), Hoboken, NJ: John Wiley & Sons.

Little, R. J. A., and Smith, P. J. (1987), "Editing and Imputation for Quantitative Survey Data," *Journal of the American Statistical Association*, 82, 58–68.

MacEachern, S. N., and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–238.

Manrique-Vallier, D., and Reiter, J. P. (2014), "Bayesian Estimation of Discrete Multivariate Latent Structure Models with Structural Zeros," *Journal of Computational and Graphical Statistics*, 23, 1061–1079.

Meng, X.-L., and Zaslavsky, A. M. (2002), "Single Observation Unbiased Priors," *The Annals of Statistics*, 30, 1345–1375.

Norberg, A. (2009), "Editing at Statistics Sweden – Yesterday, Today and Tomorrow," in *Modernisation of Statistics Production 2009*, Sockholm, Sweden.

O'Malley, A. J., and Zaslavsky, A. M. (2008), "Domain-Level Covariance Analysis for Multilevel Survey Data With Structured Nonresponse," *Journal of the American Statistical Association*, 103, 1405–1418.

Pannekoek, J., and De Waal, T. (2005), "Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the EUREDIT Project," *Journal of Official Statistics*, 21, 257–286.

Pannekoek, J., Scholtus, S., and Van der Loo, M. (2013), "Automated and Manual Data Editing: A View on Process Design and Methodology," *Journal of Official Statistics*, 29, 511–537.

Parker, J. D., and Schenker, N. (2007), "Multiple Imputation for National Public-Use Datasets And Its Possible Application for Gestational Age in United States Natality Files," *Paediatric and Perinatal Epidemiology*, 21, 97–105.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Survey Methodology*, 27, 85–95.

Riera-Ledesma, J., and Salazar-González, J.-J. (2007), "A Branch-and-Cut Algorithm for the Continuous Error Localization Problem in Data Cleaning," *Computers and Operations Research*, 34, 2790–2804.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

——— (1987), *Multiple Imputation for Nonresponse in Surveys*, Hoboken, NJ: John Wiley & Sons.

Sande, G. (1979), "Numerical Edit and Imputation," in *Proceedings of the 42nd Session of the International Statistical Institute*, Manila, Philippines.

Schiopu-Kratina, I., and Kovar, J. G. (1989), "Use of Chernikova's Algorithm in the Generalized Edit and Imputation System," Methodology Branch Working Paper BSMD 89-001E, Statistics Canada.

Scholtus, S. (2009), "Automatic Correction of Simple Typing Errors in Numerical Data With Balance Edits," Discussion Paper 09046, Statistics Netherlands, The Hague/Heerlen.

——— (2011), "Algorithms for Correcting Sign Errors and Rounding Errors in Business Survey Data," *Journal of Official Statistics*, 27, 467–490.

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

Suchard, M. A., Wang, Q., Chan, C., Frelinger, J., Cron, A., and West, M. (2010), "Understanding GPU Programming for Statistical Computation: Studies in Massively Parallel Massive Mixtures," *Journal of Computational and Graphical Statistics*, 19, 419–438.

Tempelman, C. (2007), "Imputation of Restricted Data," Ph. D. dissertation, University of Groningen.

Thompson, K. J., Fagan, J. T., Yarbrough, B. L., and Hambric, D. L. (2004), "Using a Quadratic Programming Approach to Solve Simultaneous Ratio and Balance Edit Problems," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 4485–4490.

United Nations (2006), *Statistical Data Editing: Volume No. 3, Impact on Data Quality*, Geneva, Switzerland: United Nations.

Van der Loo, M., De Jonge, E., and Scholtus, S. (2011), "Correction of Rounding, Typing, and Sign Errors With the Deducorrect Package," Discussion Paper 201119, Statistcs Netherlands, Hague/Heerlen.

Winkler, W. E., and Chen, B.-C. (2002), "Extending the Fellegi Holt Model of Statistical Data Editing," Research Report RRS2002/02, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

Winkler, W. E., and Draper, L. R. (1996), "Application of the SPEER Edit System," Research Report RR96/02, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

Woo, M. J., Reiter, J. P., Oganian, A., and Karr, A. F. (2009), "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation," *Journal of Privacy and Confidentiality*, 1, 111–124.

Supplementary Materials: Appendices

Simultaneous Edit-Imputation for Continuous Microdata

In Section APPENDIX A, we include the steps for the MCMC algorithm for the Bayesian editing model. In Section APPENDIX B, we describe methodology for checking whether or not a proposed error localization has a feasible solution. In Section APPENDIX C, we specify the parameters used in the simulation study. In Section APPENDIX D, we summarize the fitting procedures for the other editing methods used in the simulation study. In Section APPENDIX E, we present an evaluation of the quality of the edited data in the simulation study via a propensity score utility measure. In Section APPENDIX F, we present results from two additional simulation studies where A_i depends on the values of $X_i^{\mathcal{C}_i}$ or $X_i^{\mathcal{E}_i}$, respectively. In Section APPENDIX G, we present results of a simulation study involving outlier detection methods. Finally, in Section APPENDIX H we provide more information on the edit rules for the metalworking machinery manufacturing industry we use in the 2007 Census of Manufactures application.

APPENDIX A. MCMC STEPS FOR BAYESIAN EDITING

We estimate the Bayesian editing model from Section 3.2 in the main text (graphically represented in Figure A.1) using an MCMC sampler. Here, any $\boldsymbol{y}_i$ that passes all edits is treated as a true value, i.e., $s_{ij} = 0$ and $x_{ij} = y_{ij}$ for all j. We also set $s_{ij} = 0$ for any other value known to be correct (e.g., from manual edits), and we fix $s_{ij} = 1$ when y_{ij} is missing or

violates range restrictions. For the remaining cases, we make s_{ij} missing. Each iteration t of the MCMC involves the following steps.

Step 1. For each edit-failing record i, update $(\boldsymbol{x}_i, \boldsymbol{s}_i)$.

1. Propose a feasible $\boldsymbol{s}_i'$ from neighbors of $\boldsymbol{s}^{(t-1)}$.

   (a) Define the set comprising of the all-but-total variables and total variables that are involved in the failed balance edits by $NT^* = NT \cup \{T_l : l \notin \mathcal{B}_{\text{pass}}\}$. Enumerate all feasible proposals $\{\boldsymbol{s}_i' = (s_{i1}', \ldots, s_{ip}')\}$ for each of the following procedures.

      i. Birth procedure (`B`)

      Randomly choose a variable j' from $\{j : s_{ij}^{(t-1)} = 0, j \in NT^*\}$, and let $s_{ij'}' = 1$. For variables $j \in NT^* \setminus \{j'\}$, let $s_{ij}' = s_{ij}^{(t-1)}$. For each $l \in \mathcal{B}_{\text{pass}}$, let $s_{iT_l}' = \max_{j^* \in B_l} s_{ij^*}'$.

      ii. Swap procedure (`S`)

      Randomly choose a variable j_1' from $\{j : s_{ij}^{(t-1)} = 0, j \in NT^*\}$, and let $s_{ij_1'}' = 1$. Randomly choose a variable j_2' from $\{j : s_{ij}^{(t-1)} = 1, j \in NT^*\}$, and let $s_{ij_2'}' = 0$. For variables $j \in NT^* \setminus \{j_1', j_2'\}$, let $s_{ij}' = s_{ij}^{(t-1)}$. For each $l \in \mathcal{B}_{\text{pass}}$, let $s_{iT_l}' = \max_{j^* \in B_l} s_{ij^*}'$.

      iii. Death procedure (`D`)

      Randomly choose a variable j' from $\{j : s_{ij}^{(t-1)} = 1, j \in NT^*\}$, and let $s_{ij'}' = 0$. For variables $j \in NT^* \setminus \{j'\}$, let $s_{ij}' = s_{ij}^{(t-1)}$. For each $l \in \mathcal{B}_{\text{pass}}$, let $s_{iT_l}' = \max_{j^* \in B_l} s_{ij^*}'$.

   (b) For each procedure, count the number of feasible $\boldsymbol{s}_i'$ denoted by $n_{\texttt{proc}}$, for $\texttt{proc} = \texttt{B}, \texttt{S}, \texttt{D}$. See Appendix B for how to check if $\boldsymbol{s}_i'$ is feasible.

   (c) Let g_a' denote the inverse of the number of "available" procedures for which $n_{\texttt{proc}} >$

0. Randomly choose a procedure $\texttt{proc}'$ among the available procedures with probability g'_a.

(d) Randomly choose a $\boldsymbol{s}'_i$ from the set of the feasible proposals for the chosen procedure $\texttt{proc}'$ with equal probability $g'_b = 1/n_{\texttt{proc}'}$. The proposal distribution of $\boldsymbol{s}'_i$ is $g(\boldsymbol{s}'_i|\boldsymbol{s}^{(t-1)}) = g'_a g'_b$.

2. Propose $\boldsymbol{x}'_i$ given $\boldsymbol{s}'_i$.

(a) For unflagged variables $\{j : s'_{ij} = 0\}$, let $x'_{ij} = y_{ij}$.

(b) Find the set of "free" variables $\mathcal{F}_i$, to be drawn from a joint distribution, by the following rule.

- All flagged variables not involved in any balance edit are free variables, i.e., $\mathcal{F}_i \supset \{j : s'_{ij} = 1, j \notin B_l, j \neq T_l, l = 1, \ldots, q\}$.
- For each balance edit l, there are $(n'_{B_l} + s'_{iT_l} - 1)$-free variables where $n'_{B_l} = \sum_{j \in B_l} s'_{ij}$ is the number of flagged component variables.

  (i) If the total variable involved in balance edit j is flagged, all flagged component variables involved in l are free variables, i.e., $\mathcal{F}_i \supset \{j : s'_{ij} = 1, j \in B_l\}$ if $s'_{iT_l} = 1$.

  (ii) If the total variable involved in balance edit j is unflagged, the first $(n'_{B_l} - 1)$ component variables involved in l are free variables, i.e., $\mathcal{F}_i \supset \{j_m : s'_{ij_m} = 1, j_m \in B_l, m = 1, \ldots, n'_{B_l} - 1\}$ if $s'_{iT_l} = 0$. For example, if $(s'_{i1}, s'_{i2}, s'_{i3}) = (0, 1, 1)$ and the balance edit is $x_{i1} = x_{i2} + x_{i3}$, the second variable is a free variable whose value x'_{i2} will be drawn from a joint distribution.

Define the set of unflagged component variables by $UF_i = \{j : s'_{ij} = 0, j \in NT\}$ and the set of other component variables by $R_i = \{j : s'_{ij} = 1, j \notin \mathcal{F}_i, j \in NT\}$.

(c) Draw proposed values for the free variables $\boldsymbol{x}'_{i,\mathcal{F}} = \{x'_{ij} : j \in \mathcal{F}_i\}$ from a conditional normal distribution given observed values for unflagged variables $\boldsymbol{y}^{UF}_{i,NT} = \{y_{ij} : s'_{ij} = 0, j \neq T_l, l = 1, \ldots, q\}$. For the parameters of the conditional normal distribution, $\boldsymbol{\mu}^{(t-1)}_{z_i^{(t-1)}}$ and $\boldsymbol{\Sigma}^{(t-1)}_{z_i^{(t-1)}}$ and $\boldsymbol{y}_i$ are partitioned as

$$\boldsymbol{\mu}^{(t-1)}_{z_i^{(t-1)}} = \begin{pmatrix} \boldsymbol{\mu}_{\mathcal{F}} \\ \boldsymbol{\mu}_U \\ \boldsymbol{\mu}_R \end{pmatrix}, \boldsymbol{\Sigma}^{(t-1)}_{z_i^{(t-1)}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathcal{F}} & \boldsymbol{\Sigma}_{\mathcal{F},U} & \boldsymbol{\Sigma}_{\mathcal{F},R} \\ \boldsymbol{\Sigma}_{U,\mathcal{F}} & \boldsymbol{\Sigma}_U & \boldsymbol{\Sigma}_{U,R} \\ \boldsymbol{\Sigma}_{R,\mathcal{F}} & \boldsymbol{\Sigma}_{R,U} & \boldsymbol{\Sigma}_R \end{bmatrix}$$

with the sorted rows (and columns) corresponding to the order of variables in $\mathcal{F}_i$, $UF_{i,NT}$ and $R_{i,NT}$, respectively. Then, we randomly propose the value of $\boldsymbol{x}'_{i,\mathcal{F}}$ from

$$\log \boldsymbol{x}'_{i,\mathcal{F}} \sim \mathrm{N}\left(\boldsymbol{\mu}'_i, \boldsymbol{\Sigma}'_i\right).$$

where $\boldsymbol{\mu}'_i = \boldsymbol{\mu}_{\mathcal{F}} + \boldsymbol{\Sigma}_{\mathcal{F},U}\boldsymbol{\Sigma}_U^{-1}(\boldsymbol{y}^U_{i,NT} - \boldsymbol{\mu}_U)$ and $\boldsymbol{\Sigma}'_i = \boldsymbol{\Sigma}_{\mathcal{F}} - \boldsymbol{\Sigma}_{\mathcal{F},U}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{\Sigma}_{U,\mathcal{F}}$.

(d) The remaining flagged variables $\{j : s'_{ij} = 1, j \notin \mathcal{F}_i\}$ are calculated by the balance edits $x'_{iT_l} = \sum_{j \in B_l} x'_{ij}$. For example, assume the balance edit is given as $x_1 = x_2 + x_3$. For MCMC iterations where $(s'_{i1}, s'_{i2}, s'_{i3}) = (1, 1, 0)$ is proposed, we impute a value of x_{i2} inside the feasible space and then construct $x'_{i1} = x'_{i2} + x'_{i3}$ where $x'_{i3} = y_{i3}$. For MCMC iterations where $(s'_{i1}, s'_{i2}, s'_{i3}) = (0, 1, 1)$ is proposed, we impute a value of x_{i2} inside the feasible space and then make $x'_{i3} = x'_{i1} - x'_{i2}$ where $x'_{i1} = y_{i1}$.

The proposal distribution of $\boldsymbol{x}'_i$ is expressed by

$$g(\boldsymbol{x}'_i|\boldsymbol{s}'_i) = \left[\prod_{\{j:s'_{ij}=0\}} \delta\left(x'_{ij} - y_{ij}\right)\right] \mathrm{N}\left(\log \boldsymbol{x}'_{i,\mathcal{F}}|\boldsymbol{\mu}'_i, \boldsymbol{\Sigma}'_i\right) \prod_{l=1}^{q} \delta\left(x'_{iT_l} - \sum_{j \in B_l} x'_{ij}\right).$$

3. Accept $(\boldsymbol{x}_i', \boldsymbol{s}_i')$ with the acceptance probability

$$\min\left\{1, \frac{f(\boldsymbol{y}_i|\boldsymbol{x}_i', \boldsymbol{s}_i')f(\boldsymbol{x}_i'|\boldsymbol{\theta})f(\boldsymbol{s}_i')}{f(\boldsymbol{y}_i|\boldsymbol{x}_i^{(t-1)}, \boldsymbol{s}_i^{(t-1)})f(\boldsymbol{x}_i^{(t-1)}|\boldsymbol{\theta})f(\boldsymbol{s}_i^{(t-1)})} \cdot \frac{g(\boldsymbol{x}_i^{(t-1)}|\boldsymbol{s}_i^{(t-1)})g(\boldsymbol{s}_i^{(t-1)}|\boldsymbol{s}_i')}{g(\boldsymbol{x}_i'|\boldsymbol{s}_i')g(\boldsymbol{s}_i'|\boldsymbol{s}_i^{(t-1)})}\right\}.$$

Step 2. Update $\boldsymbol{x}_i$ given $(\boldsymbol{x}_i^{(t)}, \boldsymbol{s}_i^{(t)})$ for each edit-failing record i. This step is not required to build a stationary distribution, but adding it improves the mixing and convergence properties of the Markov chains. We propose $\boldsymbol{x}_i'$ given $\boldsymbol{s}_i^{(t)}$ by again running Step 1.2. Accept $\boldsymbol{x}_i'$ with the acceptance probability

$$\min\left\{1, \frac{f(\boldsymbol{y}_i|\boldsymbol{x}_i', \boldsymbol{s}_i^{(t)})f(\boldsymbol{x}_i'|\boldsymbol{\theta})}{f(\boldsymbol{y}_i|\boldsymbol{x}_i^{(t)}, \boldsymbol{s}_i^{(t)})f(\boldsymbol{x}_i^{(t)}|\boldsymbol{\theta})} \cdot \frac{g(\boldsymbol{x}_i^{(t)}|\boldsymbol{s}_i^{(t)})}{g(\boldsymbol{x}_i'|\boldsymbol{s}_i^{(t)})}\right\}.$$

Step 3. For each $i = 1, \ldots, n$, draw $z_i^{(t)} \sim \text{Categorical}(\pi_{i1}^*, \ldots, \pi_{iK}^*)$ where

$$\pi_{ik}^* = \frac{\pi_k \text{N}(\log \boldsymbol{x}_{i,NT}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'=1}^K \pi_{k'} \text{N}(\log \boldsymbol{x}_{i,NT}|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})} \quad \text{for} \quad k = 1, \ldots, K.$$

Step 4. Jointly sample auxiliary values $(N_{\text{aug}}^{(t)}, \boldsymbol{X}_{N_{\text{aug}}-n}^{(t)}, \boldsymbol{z}_{N_{\text{aug}}-n}^{(t)})$ where $\boldsymbol{z}_{N_{\text{aug}}-n}^{(t)} = \{z_{n+1}^{(t)}, \ldots, z_{N_{\text{aug}}}^{(t)}\}$. Starting with $c_{\text{in}} = c_{\text{out}} = 0$,

1. Draw $z' \sim \text{Categorical}(\pi_1, \ldots, \pi_K)$.

2. Draw $\log \boldsymbol{x}_{NT}' \sim \text{N}(\boldsymbol{\mu}_{z'}, \boldsymbol{\Sigma}_{z'})$ and calculate total variables $\{x_{T_l}', l = 1, \ldots, q\}$ by balance edits $x_{T_l}' = \sum_{j \in B_l} x_j'$.

3. If $\boldsymbol{x}' \in \mathcal{D}$, set $c_{\text{in}} = c_{\text{in}} + 1$.
   If $\boldsymbol{x}' \notin \mathcal{D}$, set $c_{\text{out}} = c_{\text{out}} + 1$, $\boldsymbol{x}_{n+c_{\text{out}}}^{(t)} = \boldsymbol{x}'$, and $z_{n+c_{\text{out}}}^{(t)} = z'$.

4. Repeat (a) through (c) until $c_{\text{in}} = n$.

5. Let $N_{\text{aug}}^{(t)} = n + c_{\text{out}}$.

Step 5. For each $k = 1, \ldots, K$, draw $\boldsymbol{\Sigma}_k \sim \text{InverseWishart}(\zeta_k, \boldsymbol{\Phi}_k)$ and then draw $\boldsymbol{\mu}_k \sim \text{N}\left(\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k/h_k\right)$ where $N_k = \sum_{i=1}^{N_{\text{aug}}} I[z_i = k]$, $\zeta_k = N_k + \zeta_0$, $h_k = N_k + h_0$, $\boldsymbol{\mu}_k^* = (N_k \bar{\boldsymbol{x}}_k + h_0 \boldsymbol{\mu}_0)/h_k$ and $\boldsymbol{\Phi}_k = \boldsymbol{\Phi} + \boldsymbol{S}_k + (\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_0)^T/(1/N_k + 1/h_0)$. The sample mean and the sum of squared distances are calculated by $\bar{\boldsymbol{x}}_k = \sum_{\{i:z_i=k\}} \log \boldsymbol{x}_{i,NT}/N_k$ and $\boldsymbol{S}_k = \sum_{\{i:z_i=k\}} (\log \boldsymbol{x}_{i,NT} - \bar{\boldsymbol{x}}_k)(\log \boldsymbol{x}_{i,NT} - \bar{\boldsymbol{x}}_k)^T$.

Step 6. For each $k = 1, \ldots, K-1$, draw $v_k \sim \text{Beta}\left(1 + N_k, \alpha + \sum_{g>k} N_g\right)$ and let $v_K = 1$. Then calculate the mixture component weights $\pi_k = v_k \prod_{g<k}(1 - v_g)$ for $k = 1, \ldots, K$.

Step 7. For each $j = 1, \ldots, p - q$, draw $\Phi_j \sim \text{Gamma}\left(a_\Phi + \zeta_0 K/2, b_\Phi + \sum_{k=1}^K \Sigma^{-1}_{k(j,j)}/2\right)$ where $\Sigma^{-1}_{k(j,j)}$ is the j-th diagonal element of $\boldsymbol{\Sigma}_k^{-1}$.

Step 8. Draw α from $\text{Gamma}\left(a_\alpha + K - 1, b_\alpha - \log \pi_K\right)$.

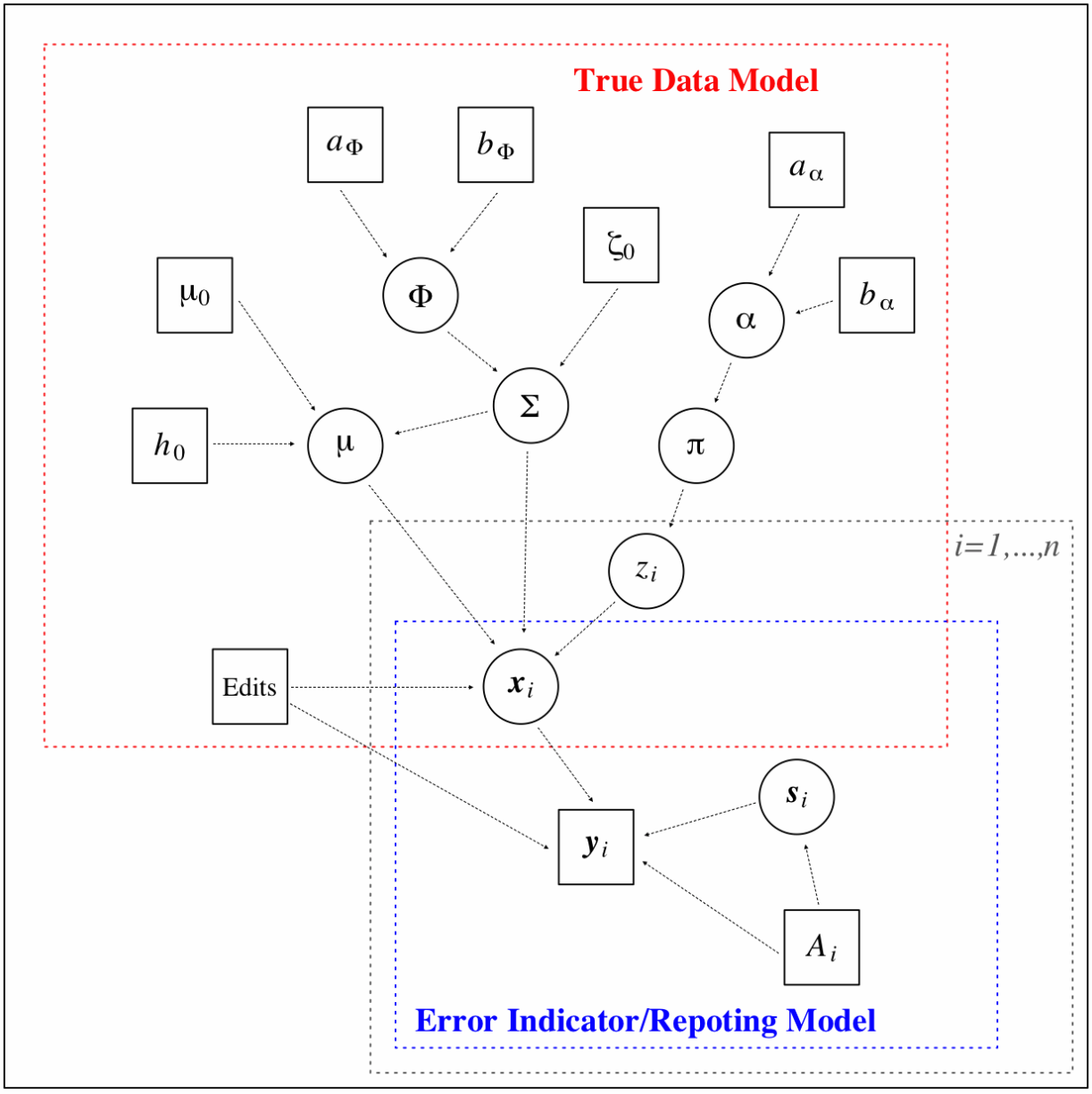

Figure A.1: Graphical representation of the Bayesian hierarchical model introduced in Section 3.2 of the main text. Rectangles represent observed data or fixed hyperparameters and circles represent unobservables.

APPENDIX B. FINDING THE SET OF FEASIBLE SOLUTIONS OF $\boldsymbol{s}_i$

In the MCMC, only proposed $\boldsymbol{s}_i$ that permit imputations satisfying all edits are acceptable, i.e, we require $\boldsymbol{s}_i \in \mathcal{S}(\boldsymbol{y}_i, A_i)$. For any $\boldsymbol{s}_i$, this condition can be checked using a sequential approach as follows.

1. **Initial checks.** Proposals must satisfy the following initial tests.
   (a) If the reported value of variable j is originally missing, the missing data item must be flagged.
   (b) If ratio edits fail, at least one of the variables subject to any ratio edit must be flagged.
   (c) If balance edit l fails, at least one of variables involved in edit l must be flagged, i.e., $\sum_{j \in B_l} s_{ij} + s_{iT_l} > 0$.
   (d) If balance edit l passes, the total variable must be flagged if one of its component variables is flagged, i.e., $s_{iT_l} = \max_{j \in B_l} s_{ij}$.

2. **Checking edit failures for unflagged items.** Assuming a record passes Step 1, this step checks whether $\boldsymbol{y}_i^{UF}$ passes edit rules by $\boldsymbol{A}^0 \boldsymbol{y}_i^{UF} \leq \boldsymbol{b}^{UF}$ where $\boldsymbol{A}^0$ and $\boldsymbol{b}^{UF}$ are the edit matrix and vector formed by "relevant" edit rules – range restrictions of unflagged variables, ratio edits between unflagged variables, and balance edits of which all component variables and the total variable are unflagged. See Kim et al. (2014) for details of how to express ratio edits and range restrictions by $\boldsymbol{A}^0$ and $\boldsymbol{b}^{UF}$. We express balance edits as two inequality constraints by adding a small threshold, i.e., $y_{iT_l} = \sum_{j \in B_l} y_{ij}$ implies that $y_{iT_l} - \epsilon \leq \sum_{j \in B_l} y_{ij} \leq y_{iT_l} + \epsilon$ for a small positive value ϵ, e.g., 0.6 in our empirical study. If $\boldsymbol{A}^0 \boldsymbol{y}_i^{UF} \leq \boldsymbol{b}^{UF}$, Step 2 is passed.

3. **Checking for a feasible region for flagged items.** Assuming a record passes Step 2, this step checks whether there exists a feasible region for imputed values of flagged variables, $\boldsymbol{x}_i^F$. Let the vector of the proposed values $\boldsymbol{x}_i$ pass edit rules if and only if $\boldsymbol{A}\boldsymbol{x}_i \leq \boldsymbol{b}$ where $\boldsymbol{A}$ and $\boldsymbol{b}$ are the matrix and vector formulated by all edit rules. After partitioning $\boldsymbol{x}_i$ as $\boldsymbol{x}_i = (\boldsymbol{y}_i^{UF}, \boldsymbol{x}_i^F)$, $\boldsymbol{A}\boldsymbol{x}_i \leq \boldsymbol{b}$ is re-expressed by

$$\begin{pmatrix} \boldsymbol{A}^0 & \boldsymbol{A}^{00} \\ \boldsymbol{A}^{10} & \boldsymbol{A}^1 \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_i^{UF} \\ \boldsymbol{x}_i^F \end{pmatrix} \leq \begin{pmatrix} \boldsymbol{b}^{UF} \\ \boldsymbol{b}^F \end{pmatrix}$$

where the sub-matrices are partitions of $\boldsymbol{A}$ with the sorted rows and columns corresponding to the order of variables in $(\boldsymbol{y}_i^{UF}, \boldsymbol{x}_i^F)$. Specifically, $\boldsymbol{A}^0$ and $\boldsymbol{b}^{UF}$ are the edit matrix and vector for $\boldsymbol{y}_i^{UF}$ used in Step 2, and $\boldsymbol{A}^{00}$ is a zero-matrix. We refer readers to Appendix of Kim et al. (2014) for an illustrative example of the matrix manipulation. After defining the edit matrices and vectors, we check whether there exists a non-zero density region of the proposed values $\boldsymbol{x}_i^F$ by solving the simplex algorithm in standard form: minimize $\boldsymbol{c}^T\boldsymbol{x}_i^F$ subject to $\boldsymbol{A}^1\boldsymbol{x}_i^F \leq \boldsymbol{b}^F - \boldsymbol{A}^{10}\boldsymbol{y}_i^{UF}$ and $\boldsymbol{x}_i^F \geq 0$ where $\boldsymbol{c}$ is an arbitrary objective function, e.g., $(1, \ldots, 1)$ in our examples. If there exists a feasible solution of the linear program, Step 3 is passed and let $\boldsymbol{s}_i \in \mathcal{S}(\boldsymbol{y}_i, \boldsymbol{A}_i)$.

For each proposed $\boldsymbol{s}_i$, the feasibility check described above does not need to derive all implied edits. In Step 3, we use the set of explicit edits, reported values $\boldsymbol{y}_i$, and the first phase of the simplex algorithm to check whether the suggested $\boldsymbol{s}_i$ generates non-null space of $\boldsymbol{x}_i^F$. Ideally, we would implement the simplex algorithm only once for each proposed possible solution, store the feasibility status in a look-up table, and reuse it when the proposal is revisited. When the number of variables involved in edits is not small, unfortunately it is not possible with most standard computers to store all visited unique solutions in memory. We thus recommend keeping in memory the feasibility status (feasible or not) for a large

but manageable number of recently visited unique proposals. For any proposed $\boldsymbol{s}_i$, we first check the in-memory look-up tables (implemented as one hash table per record) to see if we already know its status, and retrieve it directly if it is in the look-up table. If not, we run the simplex algorithm to determine feasibility and add the solution to the look-up table. For example, in the CM application we store up to 30000 solutions from each record. We also keep and update the MCMC iteration a solution is last visited. When the storage becomes full, we bump the 30% of the stored solutions with last-visited iterations in the earliest parts of the chain.

APPENDIX C. PARAMETERS OF SIMULATION STUDY

This section describes all simulated values and edit rules used for the simulation study in Section 4 of the main text. Motivated by the CM editing system, we introduce explicit edit rules consisting of the ratio edits $(L_{j,j'}, U_{j,j'})$ and balance edits shown in Table C.1. The table also displays the support of reported values $\mathcal{Y} = (\tilde{L}_1, \tilde{U}_1) \times \ldots \times (\tilde{L}_p, \tilde{U}_p)$. Note that it is natural to assume $\mathcal{X} \subset \mathcal{Y}$ in data editing for continuous data. To make $\mathcal{X}$ bounded, we set the range restrictions equal to the ranges of reported values, i.e., $(L_j, U_j) = (\tilde{L}_j, \tilde{U}_j)$ for $j = 1, \ldots, p$.

We generate a population $\boldsymbol{X}^{\text{pop}}$ by following steps: (i) generate $\boldsymbol{x}$ by sampling the all-but-total variables $\boldsymbol{x}_{NT} = (x_2, x_3, x_4, x_6, x_7, x_8)$ from a mixture of three multivariate normal distributions, $\log \boldsymbol{x}_{NT} \sim \mathrm{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ where $z \sim \text{Categorical}(0.25, 0.60, 0.15)$, (ii) set x_1 and x_5 equal to sums of their component variables, (iii) accept the generated value if $\boldsymbol{x} \in \mathcal{X}$ and reject it otherwise. Repeat steps (i)–(iii) until 1,000,000 records of $\boldsymbol{x}_i \in \mathcal{X}$ are generated. The simulation parameters used are

$$\boldsymbol{\mu}_1 = (3, 6, 5, 2, 3, 8, 7)^T, \boldsymbol{\mu}_2 = (4, 9, 8, 4, 5, 9, 7.5)^T, \boldsymbol{\mu}_3 = (7, 5, 7, 5, 5, 10, 9)^T,$$

Table C.1: Nine variables used in the simulation study of Section 4 of the main text. The left table shows the ranges of reported values $(\tilde{L}_j, \tilde{U}_j)$ and the balance edits which we introduce for illustration. The right table shows that the ratio edits $(L_{j,j'}, U_{j,j'})$ on each pair of variables. There are no explicit range restrictions.

Var.	$\tilde{L}_j$	$\tilde{U}_j$	Balance Edit
x_1	2.1	1200000	$x_1 = x_2 + x_3 + x_4$
x_2	0.1	100000	
x_3	1.0	1000000	
x_4	1.0	100000	
x_5	0.2	200000	$x_5 = x_6 + x_7$
x_6	0.1	100000	
x_7	0.1	100000	
x_8	1.0	1000000	
x_9	1.0	1000000	

$x_j/x_{j'}$	$L_{j,j'}$	$U_{j,j'}$
x_1/x_5	0.367	148.41
x_1/x_8	0.082	20.09
x_1/x_9	0.007	148.41
x_5/x_8	0.006	2.72
x_5/x_9	0.007	148.41
x_8/x_9	0.007	148.41

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.0 & 0.2 & 0.1 & -0.1 & -0.2 & 0.4 & 0.1 \\ 0.2 & 1.0 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 \\ 0.1 & 0.1 & 1.0 & -0.3 & 0.0 & 0.0 & 0.1 \\ -0.1 & 0.0 & -0.3 & 1.0 & 0.4 & -0.2 & 0.1 \\ -0.2 & 0.0 & 0.0 & 0.4 & 1.0 & -0.1 & 0.1 \\ 0.4 & 0.2 & 0.0 & -0.2 & -0.1 & 1.0 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 1.0 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.7 & -0.2 & -0.2 & 0.0 & 0.0 & 0.1 & 0.1 \\ -0.2 & 0.7 & -0.2 & 0.0 & 0.0 & 0.2 & 0.1 \\ -0.2 & -0.2 & 0.7 & 0.0 & 0.0 & 0.0 & 0.1 \\ 0.0 & 0.0 & 0.0 & 0.5 & 0.1 & -0.1 & 0.1 \\ 0.0 & 0.0 & 0.0 & 0.1 & 0.5 & -0.1 & 0.1 \\ 0.1 & 0.2 & 0.0 & -0.1 & -0.1 & 0.5 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.8 \end{bmatrix},$$

and

$$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 0.5 & 0.1 & 0.2 & -0.2 & 0.0 & 0.1 & 0.1 \\ 0.1 & 0.3 & -0.2 & 0.0 & -0.2 & 0.1 & 0.1 \\ 0.2 & -0.2 & 0.7 & 0.0 & 0.0 & 0.1 & 0.1 \\ -0.2 & 0.0 & 0.0 & 0.4 & 0.0 & 0.0 & 0.1 \\ 0.0 & -0.2 & 0.0 & 0.0 & 0.4 & 0.0 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.0 & 0.0 & 0.4 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.7 \end{bmatrix}.$$

We sample $R = 500$ independent simple random samples $\boldsymbol{X}^r$, where $r = 1, \ldots, R$, of size $n = 1000$ from $\boldsymbol{X}^{\text{pop}}$. In each $\boldsymbol{X}^r$, we randomly select 600 records to have $A_i = 0$ for which $\boldsymbol{x}_i = \boldsymbol{y}_i$, 200 records to have $A_i \in \{1, 3\}$, and 200 records to have $A_i = 2$. Within each group of 200 records, we randomly draw the simulated values of $\boldsymbol{s}_i$ and $\boldsymbol{y}_i$ by rejection sampling as follows. First, we randomly draw $s_{ij} \sim \text{Bernoulli}(\psi_{s_j})$ for $j = 1, \ldots, p$ where $(\psi_{s_1}, \ldots, \psi_{s_9}) = (0.41, 0.21, 0.08, 0.29, 0.31, 0.11, 0.09, 0.32, 0.29)$. Second, we randomly draw values of y_{ij} whenever $s_{ij} = 1$ from the appropriate uniform distribution with the limits $(\tilde{L}_j, \tilde{U}_j)$ shown in Table C.1. Finally, we accept the proposed $\boldsymbol{y}_i$ when it violates edit rules in accord with the corresponding A_i, i.e., $\boldsymbol{y}_i$ with $A_i \in \{1, 3\}$ fails at least one balance edit and $\boldsymbol{y}_i$ with $A_i = 2$ passes all balance edits but fails at least one inequality constraint. Specifically for records with $A_i = 2$, we propose values of y_{ij} for $s_{ij} = 1$ and $j \notin \{1, 5\}$ from the uniform distribution, calculate the total variables $y_{i1} = y_{i2} + y_{i3} + y_{i4}$ and $y_{i5} = y_{i6} + y_{i7}$, and check if $\boldsymbol{y}_i$ violates inequality constraints. In our process of sampling the simulated values by rejection sampling, typically around 79% of $(\boldsymbol{s}_i, \boldsymbol{y}_i)$ with $A_i \in \{1, 3\}$ are accepted and around 46% of $(\boldsymbol{s}_i, \boldsymbol{y}_i)$ with $A_i = 2$ are accepted.

APPENDIX D. MCMC STEPS FOR OTHER EDITING METHODS

To implement the editing methods in Section 4 of the main text, we modify the MCMC steps for `BE` described in Appendix A as follows.

`FH`: For each i, determine the error indicators $\boldsymbol{s}_i$ under MWFI criterion. We set the assumed reliability weight of item j by $w_j = 1 - \sum_{i=1}^{n} s_{ij}^*/n$ where s_{ij}^* is the simulated (true) $\boldsymbol{s}_{ij}$ used to generate the simulated reported dataset. Given the fixed value of $\boldsymbol{s}_i$, we implement MCMC Steps 2–8 in Appendix A.

`BE-min`: The feasible region of $\boldsymbol{s}$ is restricted to $\mathcal{S}_{\text{FH}}(\boldsymbol{y}_i, A_i)$ which is the set of $\boldsymbol{s}_i' \in \mathcal{S}(\boldsymbol{y}_i, A_i)$ such that $\sum_j s_{ij}' \leq \sum_j s_{ij}$ for all other $\boldsymbol{s}_i \in \mathcal{S}(\boldsymbol{y}_i, A_i)$. Then, we run the MCMC in Appendix A with replacing $\mathcal{S}(\boldsymbol{y}_i, A_i)$ with $\mathcal{S}_{\text{FH}}(\boldsymbol{y}_i, A_i)$. In Step 1.1.(a), we only need the swap procedure because $\sum_j \boldsymbol{s}_{ij}'$ is fixed to $\sum_j \boldsymbol{s}_{ij}^{(t-1)}$.

`AAI`: For each i, determine $\boldsymbol{s}_i$ to flag all active items, which have conflict with other data items. That is, let $s_{ij} = 1$ for variables j that are involved in failed edits (including balance edits, ratio edits and range restrictions). Additionally, we set $s_{ij} = 1$ for all $j \in B_l$ if $s_{iT_l} = 1$. Given the fixed value of $\boldsymbol{s}_i$, we implement MCMC Steps 2–8 in Appendix A.

`BE-sgl`: Implement the MCMC steps in Appendix A with modifying steps as follows:

· In Step 1, omit subscript $z_i^{(t-1)}$ from $\boldsymbol{\mu}_{z_i^{(t-1)}}^{(t-1)}$ and $\boldsymbol{\Sigma}_{z_i^{(t-1)}}^{(t-1)}$.

· Omit Steps 3, 6, 7 and 8.

· Modify Step 4 as follows: Starting with $c_{\text{in}} = c_{\text{out}} = 0$,

(a) Draw $\log \boldsymbol{x}'_{NT} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and calculate total variables $\{x'_{T_l}, l = 1, \ldots, q\}$ by balance edits $x'_{T_l} = \sum_{j \in B_l} x'_j$.

(b) If $\boldsymbol{x}' \in \mathcal{D}$, set $c_{\text{in}} = c_{\text{in}} + 1$.
If $\boldsymbol{x}' \notin \mathcal{D}$, set $c_{\text{out}} = c_{\text{out}} + 1$, $\boldsymbol{x}^{(t)}_{n+c_{\text{out}}} = \boldsymbol{x}'$.

(c) Repeat (a) through (c) until $c_{\text{in}} = n$.

(d) Let $N^{(t)}_{\text{aug}} = n + c_{\text{out}}$.

· Modify Step 5 as follows.

Draw $\boldsymbol{\Sigma} \sim \text{InverseWishart}(\zeta^*, \boldsymbol{\Phi}^*)$ and then draw $\boldsymbol{\mu} \sim \text{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}/h^*)$ where $\zeta^* = N_{\text{aug}} + \zeta_0$, $h^* = N_{\text{aug}} + h_0$, $\boldsymbol{\mu}^* = (N_{\text{aug}}\bar{\boldsymbol{x}}^* + h_0\boldsymbol{\mu}_0)/h^*$, $\boldsymbol{\Phi}^* = \boldsymbol{\Phi} + \boldsymbol{S}^* + (\boldsymbol{\mu}^* - \boldsymbol{\mu}_0)(\boldsymbol{\mu}^* - \boldsymbol{\mu}_0)^T/(1/N_{\text{aug}} + 1/h_0)$. The sample mean and the sum of squared distances are calculated by $\bar{\boldsymbol{x}}^* = \sum_{i=1}^{N_{\text{aug}}} \log \boldsymbol{x}_{i,NT}/N_{\text{aug}}$ and $\boldsymbol{S}^* = \sum_{i=1}^{N_{\text{aug}}} (\log \boldsymbol{x}_{i,NT} - \bar{\boldsymbol{x}}^*)(\log \boldsymbol{x}_{i,NT} - \bar{\boldsymbol{x}}^*)^T$. In our simulation study, we put weak priors for $\boldsymbol{\Phi} = diag(\Phi_1, \ldots, \Phi_{p-q})$ where $\Phi_j = 0.01$ for $j = 1, \ldots, p - q$.

`TrueS`

Implement MCMC Steps 2–8 in Appendix A with each $\boldsymbol{s}_i$ fixed at the simulated values $\boldsymbol{s}^*_i$.

APPENDIX E. EVALUATIONS OF PROPENSITY SCORE UTILITY MEASURE

In the simulation study in Section 4 of the main text, we also evaluated data quality using the propensity score metric from the literature on statistical disclosure limitation (Woo et al. 2009), which determines how well one can discriminate $\boldsymbol{X}^{r(m)}$ from $\boldsymbol{X}^r$. Inability to discriminate suggests similar distributions, which means high quality of $\boldsymbol{X}^{r(m)}$. The propensity score for some $\boldsymbol{X}^{r(m)}$ and $\boldsymbol{X}^r$ is calculated as follows.

Table E.1: Summary of propensity score utility measures PS_r (times 100) across the 500 simulations.

	TrueS	BE	FH	BE-min	AAI	BE-sgl
Mean	3.56	6.40	20.94	16.23	10.06	12.00
S.E.	0.02	0.03	0.10	0.13	0.04	0.04

1. Concatenate $\boldsymbol{X}^r$ and $\boldsymbol{X}^{r(m)}$, and add an indicator variable with values equal to zero for all records in $\boldsymbol{X}^r$ and equal to one for all records in $\boldsymbol{X}^{r(m)}$.

2. Using the concatenated data, estimate the logistic regression of the indicator variable on all $p = 9$ variables (after log transformations), including main effects and all interactions up to third order. Thus, we estimate the logistic regression function,

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{a=1}^{9}\beta_a \log x_{ia} + \sum_{a,b}\beta_{ab}\log x_{ia}\log x_{ib} + \sum_{a,b,c}\beta_{abc}\log x_{ia}\log x_{ib}\log x_{ic}. \tag{E.1}$$

3. For $i = 1, \ldots, 2n$, we compute the set of predicted probabilities $\hat{p}_i$ using the MLEs of the coefficients in (E.1).

4. The propensity score utility measure for $\boldsymbol{X}^{r(m)}$ is defined as

$$\text{PS}^{r(m)} = \frac{1}{2n}\sum_{i=1}^{2n}\left(\hat{p}_i - \frac{1}{2}\right)^2.$$

For each $\boldsymbol{X}^r$, we compute $\text{PS}^r = \sum_{m=1}^{M}\text{PS}^{r(m)}/M$. Table E.1 shows that the propensity score of BE is much lower than other methods (about a third of that of FH), implying that the multivariate distribution in the edited data from BE is more similar to the multivariate distribution of $\boldsymbol{X}^r$ than those from other editing methods.

APPENDIX F. SIMULATION STUDIES WHERE A_i DEPENDS ON $\boldsymbol{x}_i^{\mathcal{C}_i}$ OR $\boldsymbol{x}_i^{\mathcal{E}_i}$.

In this section, we apply the Bayesian editing to simulation examples where the data generating distribution for A_i depends on $\boldsymbol{x}_i$. We modify the simulation design of the main text as follows. We use the population $\boldsymbol{X}^{\text{pop}}$ and its $R = 500$ samples generated in Section 4. In each $\boldsymbol{X}^r$, the values of A_i for $i = 1, \ldots, n$ are randomly drawn with probabilities $P(A_i = 0) = 1 - 1/[1 + \exp(-u_i)]$ and $P(A_i = 1) = P(A_i = 2) = 1/2[1 + \exp(-u_i)]$ where $u_i = 17 - 2x_{i,8}$.

We assume two scenarios where the data generating distribution for A_i depends on one variable known to be correct for all records $\boldsymbol{x}_i^{\mathcal{C}_i}$ (FAR scenario) and on a variable that is subject to reporting errors $\boldsymbol{x}_i^{\mathcal{E}_i}$ (not FAR scenario). For the FAR scenario, for records with $A_i > 0$, we generate the simulated values of $\boldsymbol{s}_i$ by randomly drawing $s_{ij} \sim \text{Bernoulli}(\psi_{s_j})$ for $j = 1, \ldots, p$ until the generated $\boldsymbol{s}_i$ is in the feasible region $\mathcal{S}(\boldsymbol{y}_i, A_i)$ where $(\psi_{s_1}, \ldots, \psi_{s_9}) = (0.41, 0.21, 0.08, 0.29, 0.31, 0.11, 0.09, 0.00, 0.29)$. Thus, variable $j = 8$ is assumed known to be correct and has $s_{i8} = 0$ for all i. For the not FAR scenario, we generate the simulated values of $\boldsymbol{s}_i$ with $(\psi_{s_1}, \ldots, \psi_{s_9}) = (0.41, 0.21, 0.08, 0.29, 0.31,\ 0.11, 0.09, 0.32, 0.29)$. Thus, variable $j = 8$ may or may not be known to be correct. For records with $A_i = 0$, set $\boldsymbol{y}_i = \boldsymbol{x}_i$. For records with $A_i > 0$, models in (11) and (12) in the main text are used to generate $\boldsymbol{y}_i$ corresponding to the given values of $\boldsymbol{s}_i$, for both of scenarios.

Tables F.1 and F.2 summarize the simulation results of the FAR scenario. The overall results are similar to Table 1 of the main text. `BE` tends to result in the smallest biases and root mean squared errors (RMSE) among all the editing methods, and its 95% coverage rates are close to the nominal values.

Tables F.3 and F.4 summarize the results from the not FAR scenario. Because none of

Table F.1: Summaries of the estimators of population mean across $R = 500$ simulations in the FAR scenario – the data generating distribution for A_i depends on x_{i8} which is assumed known to be correct. The columns labeled $\boldsymbol{X}^r$ and $\boldsymbol{X}^{r,\text{pass}}$ display results based on the true data and the edit-passing records only, respectively. The column labeled `TrueS` displays the results using the model in Section 3 with each s_{ij} fixed at the true s^*_{ij}.

				Editing methods				
	$\boldsymbol{X}^r$	$\boldsymbol{X}^{r,\text{pass}}$	`TrueS`	`BE`	`FH`	`BE-min`	`AAI`	`BE-sgl`
relBias (×100)								
$\bar{X}_1$	0.1	15.1	0.2	-0.1	3.4	3.2	1.5	2.3
$\bar{X}_2$	0.6	44.2	1.4	0.2	95.3	70.6	5.3	0.0
$\bar{X}_3$	0.2	15.6	0.2	0.0	-0.6	0.2	0.7	4.0
$\bar{X}_4$	0.0	11.6	0.1	-0.3	4.1	3.8	2.7	-1.2
$\bar{X}_5$	-0.1	11.0	-0.2	-1.3	38.5	21.6	5.1	-2.8
$\bar{X}_6$	-0.1	15.5	-0.1	-1.0	86.0	38.7	5.6	-1.8
$\bar{X}_7$	-0.1	9.0	-0.3	-1.5	17.7	14.0	4.9	-3.3
$\bar{X}_8$	0.4	35.5	0.4	0.4	0.4	0.4	0.4	0.4
$\bar{X}_9$	0.0	23.6	0.3	0.0	36.9	37.9	6.6	-0.3
relRMSE (×100)								
$\bar{X}_1$	2.7	15.5	2.7	2.8	4.6	4.4	3.4	3.7
$\bar{X}_2$	8.1	45.9	8.3	8.1	100.6	76.6	10.6	8.1
$\bar{X}_3$	3.5	16.3	3.5	3.6	3.7	3.7	4.1	5.5
$\bar{X}_4$	3.7	12.6	3.8	4.1	6.3	6.1	5.5	4.4
$\bar{X}_5$	2.4	11.4	2.4	2.8	40.8	23.3	6.1	3.8
$\bar{X}_6$	3.3	16.3	3.3	3.6	92.8	44.1	7.5	4.0
$\bar{X}_7$	2.8	9.7	2.8	3.3	21.0	16.8	6.4	4.4
$\bar{X}_8$	2.8	35.7	2.8	2.8	2.8	2.8	2.8	2.8
$\bar{X}_9$	4.5	24.5	4.6	4.7	40.5	41.9	9.1	4.7
95% CI Coverage								
$\bar{X}_1$	95.2	0.0	95.2	94.2	82.4	83.0	93.6	90.8
$\bar{X}_2$	93.0	2.6	94.8	94.0	2.0	17.0	94.8	93.6
$\bar{X}_3$	94.4	9.0	94.4	94.2	94.4	94.8	94.0	87.8
$\bar{X}_4$	93.4	31.4	93.6	94.2	90.8	90.0	91.8	92.6
$\bar{X}_5$	93.8	2.0	94.2	92.0	1.4	22.8	82.4	78.2
$\bar{X}_6$	94.8	5.0	95.0	92.8	6.4	63.4	92.6	89.2
$\bar{X}_7$	94.8	24.2	95.0	92.2	73.2	78.2	88.8	78.6
$\bar{X}_8$	95.0	0.0	95.0	95.0	95.0	95.0	95.0	95.0
$\bar{X}_9$	95.6	4.0	95.2	96.0	36.6	37.6	89.0	96.2

Table F.2: Summaries of the estimators of regression coefficients across $R = 500$ simulations in the FAR scenario – the data generating distribution for A_i depends on x_{i8} which is assumed known to be correct. The columns labeled $\boldsymbol{X}^r$ and $\boldsymbol{X}^{r,\text{pass}}$ display results based on the true data and the edit-passing records only, respectively. The column labeled `TrueS` displays the results using the model in Section 3 with each s_{ij} fixed at the true s^*_{ij}.

				Editing methods				
	$\boldsymbol{X}^r$	$\boldsymbol{X}^{r,\text{pass}}$	`TrueS`	`BE`	`FH`	`BE-min`	`AAI`	`BE-sgl`
relBias (×100)								
β_0	0.2	-27.4	0.7	4.3	-16.8	-14.8	-9.8	2.3
β_1	-0.8	-26.9	-0.7	-7.6	29.5	17.7	9.2	-18.7
β_2	0.0	22.1	0.1	3.3	-4.9	10.8	5.5	12.8
β_3	0.2	24.6	-0.3	-1.8	2.2	-3.3	0.0	0.1
relRMSE (×100)								
β_0	8.0	32.3	9.6	12.2	20.6	19.0	16.8	11.8
β_1	13.7	32.5	15.3	18.6	34.2	25.0	23.4	24.4
β_2	9.9	26.3	11.3	12.9	14.3	18.1	16.1	17.9
β_3	9.2	28.2	10.3	11.4	12.2	12.6	12.1	11.0
95% CI Coverage								
β_0	94.8	61.0	94.8	94.0	65.6	70.4	86.2	94.2
β_1	93.2	65.6	93.6	92.6	58.0	81.0	92.0	79.4
β_2	92.6	61.4	93.6	92.2	86.6	81.4	93.0	80.6
β_3	94.4	55.8	94.0	94.8	94.4	93.4	93.8	94.0

the editing procedures account for the nonignorable selection of faulty data, estimators of $\bar{X}_8$ are biased for all editing methods. Nonetheless, `BE` still tends to perform best among the editing methods.

Table F.3: Summaries of the estimators of population mean across $R = 500$ simulations in the not FAR scenario – the data generating distribution for A_i depends on x_{i8} which is subject to reporting errors. The columns labeled $\boldsymbol{X}^r$ and $\boldsymbol{X}^{r,\text{pass}}$ display results based on the true data and the edit-passing records only, respectively. The column labeled `TrueS` displays the results using the model in Section 3 with each s_{ij} fixed at the true s^*_{ij}.

				Editing methods				
	$\boldsymbol{X}^r$	$\boldsymbol{X}^{r,\text{pass}}$	`TrueS`	`BE`	`FH`	`BE-min`	`AAI`	`BE-sgl`
relBias (×100)								
$\bar{X}_1$	0.1	15.1	0.4	1.6	8.4	6.1	7.8	6.1
$\bar{X}_2$	0.6	44.2	1.9	4.3	111.3	87.5	21.6	10.8
$\bar{X}_3$	0.2	15.6	0.4	1.8	4.2	3.2	7.8	7.5
$\bar{X}_4$	0.0	11.6	0.3	1.1	8.5	5.5	6.7	2.8
$\bar{X}_5$	-0.1	11.0	-0.1	-0.9	140.7	62.7	8.2	5.4
$\bar{X}_6$	-0.1	15.5	-0.1	-0.4	292.2	118.5	10.2	18.6
$\bar{X}_7$	-0.1	9.0	-0.2	-1.1	74.3	38.3	7.3	-0.4
$\bar{X}_8$	0.4	35.5	5.0	9.4	15.3	14.4	18.0	13.6
$\bar{X}_9$	0.0	23.6	0.4	1.6	40.8	40.2	11.6	5.5
relRMSE (×100)								
$\bar{X}_1$	2.7	15.5	2.7	3.3	9.5	7.4	8.4	7.0
$\bar{X}_2$	8.1	45.9	8.5	9.9	116.9	94.9	24.1	17.0
$\bar{X}_3$	3.5	16.3	3.6	4.1	7.4	6.3	8.9	8.8
$\bar{X}_4$	3.7	12.6	3.9	4.3	10.2	7.5	8.2	5.2
$\bar{X}_5$	2.4	11.4	2.4	2.8	146.7	67.8	9.0	11.9
$\bar{X}_6$	3.3	16.3	3.4	4.0	312.3	136.5	12.0	36.3
$\bar{X}_7$	2.8	9.7	2.8	3.3	82.9	45.2	8.3	7.0
$\bar{X}_8$	2.8	35.7	5.8	9.9	16.4	15.6	18.3	14.1
$\bar{X}_9$	4.5	24.5	4.6	5.1	45.4	45.4	13.5	9.8
95% CI Coverage								
$\bar{X}_1$	95.2	0.0	95.8	90.8	52.8	74.2	36.8	62.2
$\bar{X}_2$	93.0	2.6	94.8	95.2	1.4	15.2	59.8	95.2
$\bar{X}_3$	94.4	9.0	94.6	93.0	96.0	96.4	61.8	68.4
$\bar{X}_4$	93.4	31.4	94.6	92.8	73.0	85.2	73.2	93.0
$\bar{X}_5$	93.8	2.0	94.8	93.4	0.0	19.4	61.4	95.8
$\bar{X}_6$	94.8	5.0	94.6	94.4	4.6	68.8	75.2	97.0
$\bar{X}_7$	94.8	24.2	95.0	92.4	34.8	78.8	79.2	89.2
$\bar{X}_8$	95.0	0.0	59.0	12.2	19.6	24.2	0.0	2.8
$\bar{X}_9$	95.6	4.0	95.6	95.6	44.6	41.0	65.6	95.2

Table F.4: Summaries of the estimators of regression coefficients across $R = 500$ simulations in the not FAR scenario – the data generating distribution for A_i depends on x_{i8} which is subject to reporting errors. The columns labeled $\boldsymbol{X}^r$ and $\boldsymbol{X}^{r,\text{pass}}$ display results based on the true data and the edit-passing records only, respectively. The column labeled `TrueS` displays the results using the model in Section 3 with each s_{ij} fixed at the true s^*_{ij}.

				Editing methods				
	$\boldsymbol{X}^r$	$\boldsymbol{X}^{r,\text{pass}}$	`TrueS`	`BE`	`FH`	`BE-min`	`AAI`	`BE-sgl`
relBias (×100)								
β_0	0.2	-27.4	-0.4	-2.8	-16.8	-7.5	-21.1	-13.6
β_1	-0.8	-26.9	-0.9	-8.8	50.7	32.2	0.8	-14.4
β_2	0.0	22.1	-0.1	4.6	-34.3	-7.2	9.9	13.6
β_3	0.2	24.6	0.4	3.6	8.3	-7.1	10.8	10.6
relRMSE (×100)								
β_0	8.0	32.3	10.0	12.2	21.7	15.0	26.5	19.1
β_1	13.7	32.5	15.2	19.1	53.8	36.8	20.6	21.5
β_2	9.9	26.3	11.0	13.4	37.1	17.0	17.1	18.4
β_3	9.2	28.2	11.1	12.9	16.3	15.0	17.4	16.5
95% CI Coverage								
β_0	94.8	61.0	94.8	95.8	61.6	85.2	67.0	79.6
β_1	93.2	65.6	95.4	90.4	18.6	52.8	95.4	84.8
β_2	92.6	61.4	94.8	93.4	16.4	82.8	90.8	79.0
β_3	94.4	55.8	94.2	94.8	86.4	84.4	86.0	84.4

APPENDIX G. SIMULATION STUDIES WITH OUTLIER DETECTION

Some automatic editing systems combine a minimum number of fields to impute approach with outlier detection methods (Kozak 2005). For example, the agency could flag reported values far in the tail of a univariate (or multivariate) distribution as outliers, and ensure that those values are blanked and imputed. In this section, we perform a simulation study that implements an outlier detection procedure from the Banff editing system of Statistics Canada (Banff Support Team 2007) before applying `BE-min` and `BE`.

We base the simulation study on the 500 replicated datasets described in Section 4 of the main text. For each of these 500 sets $\boldsymbol{Y}^r$, for each variable j we compute the first quartile Q_{1j}, the median Q_{2j}, and the third quartile Q_{3j} of $\log Y_{ij}^r$. We identify as outliers all y_{ij} such that $\log y_{ij} < Q_{2j} - Cd_{Q_{1j}}$ or $\log y_{ij} > Q_{2j} + Cd_{Q_{3j}}$, where $d_{Q_{1j}} = Q_{2j} - Q_{1j}$, $d_{Q_{3j}} = Q_{3j} - Q_{2j}$, and C is a pre-specified threshold. This is a univariate outlier detection scheme, finding cases that are in the tails of marginal distributions. We fix $s_{ij} = 1$ for all specified outliers, and of course $\boldsymbol{s}_i = 0$ for all cases with no outliers and no edit failures. For all other values, we let the s_{ij} be unknown and run a minimum number of fields to impute method (`BE-min`, since it outperformed `FH` in the other simulations) and the Bayesian editing method (`BE`).

Figure G.1 displays a typical result of applying outlier detection before `BE-min`. When $C = 6$, the distribution of $\log x_{i1}$ and $\log x_{i9}$ when applying outlier detection followed by `BE-min` is similar to that generated by applying `BE-min` alone, with some outlying reported values left unedited. When $C = 2$, applying outlier detection before `BE-min` results in many true values being labeled outliers and replaced with imputations. Clearly, the procedure works best when $C = 4$.

Tables G.1 and G.2 summarize the simulation results. When $C = 4$, forced editing of the

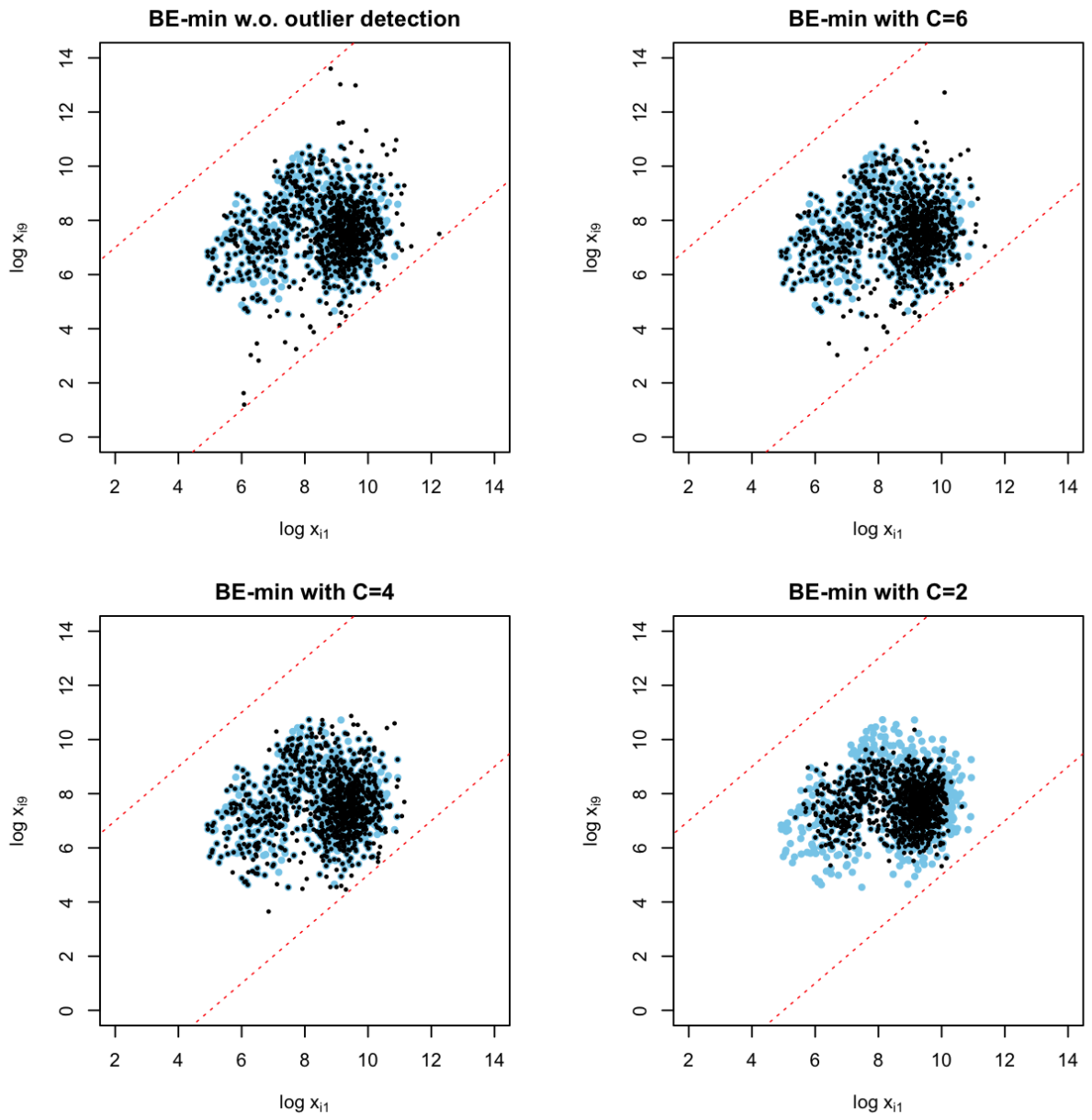

Figure G.1: Plots of $\log x_{i1}$ versus $\log x_{i9}$ when applying outlier detection followed by `BE-min`. Results are based on one randomly selected true $\boldsymbol{X}^r$ and one randomly selected dataset after applying edit-imputation to this $\boldsymbol{X}^r$. The top left panel shows the imputed values of `BE-min` without outlier detection, and other panels show `BE-min` after the outlier detection procedures with different values of C. The black dots represent values after editing and imputation, and the blue dots in background represent the true values in $\boldsymbol{X}^r$.

selected outliers improves the quality of inferences for `BE-min`. However, in this simulation, `BE` still dominates `BE-min` with outlier detection. Interestingly, setting $s_{ij} = 1$ for cases identified as outliers before running `BE` also outperforms `BE-min` with outlier detection. This suggests benefits of stochastic error localization even after outlier identification. We also ran the simulation with outlier detection on the original scale (without taking logarithms for outlier detection). Overall, the repeated sampling properties are poorer than those reported in Tables G.1 and G.2.

APPENDIX H. EDIT RULES FOR CM APPLICATION

We use data of a metalworking machinery manufacturing industry (NAICS code of 33351400) from the 2007 U.S. Economic Census, which include p = 27 variables involved in the edit rules. The industry-specific edits are provided by the Census Bureau, which were used in the SPEER system for automatic editing. The explicit edits consist of 66 ratio edits $(L_{j,j'}, U_{j,j'})$ for all possible pairs of 12 variables subject to ratio edits and six balance edits.

For confidentiality reasons, we are not able to disclose the exact values of ratio edit limits. Instead, we summarize the structure of edit rules for CM editing in Table H.1. We replace zeros in $\boldsymbol{y}_i$ with a small positive value $\omega = 0.1$ because it is necessary for ratio edits to be defined, since one cannot divide by zero (see Kovar et al. 1988). The ranges of reported values $(\tilde{L}_j, \tilde{U}_j)$ are introduced by setting $\tilde{L}_j = 0.001$ for all j and $\tilde{U}_j$ with the value either 10^4, 10^5 or 10^6 such that $\tilde{U}_j/10 < \max_i y_{ij} \leq \tilde{U}_j$. This lets all $\boldsymbol{y}_i$ be in the support $\mathcal{Y} = (\tilde{L}_1, \tilde{U}_1) \times \ldots \times (\tilde{L}_p, \tilde{U}_p)$ but not too close to the boundary of $\mathcal{Y}$. No explicit range restriction is provided in the CM data. To make $\mathcal{X}$ $(\subset \mathcal{Y})$ bounded, we set the range restrictions equal to the ranges of reported values, i.e., $(L_j, U_j) = (\tilde{L}_j, \tilde{U}_j)$ for $j = 1, \ldots, p$.

Table G.1: Summaries of the estimators of population mean across 500 simulations in outlier detection simulation. The first and second columns display results of the editing methods without an outlier detection step, which are exactly the same as the sixth and fourth columns of Table 1 of the main text, respectively. The value of C indicates the threshold parameter of the outlier detection method.

			With Outlier Detection					
			BE-min			BE		
	BE-min	BE	C=6	C=4	C=2	C=6	C=4	C=2
relBias (×100)								
$\bar{X}_1$	6.2	-0.1	4.8	2.4	-6.6	-0.1	-0.1	-5.6
$\bar{X}_2$	88.3	0.4	66.8	54.0	69.4	0.1	-8.7	-8.1
$\bar{X}_3$	3.3	0.1	1.2	-0.4	-10.5	0.2	0.5	-3.8
$\bar{X}_4$	5.5	-0.6	7.1	3.9	-4.8	-0.5	-0.7	-9.5
$\bar{X}_5$	91.3	-1.1	16.4	3.4	-5.3	-1.3	-1.9	-8.0
$\bar{X}_6$	168.6	-0.9	29.4	6.3	-1.7	-1.1	-1.4	-6.7
$\bar{X}_7$	57.4	-1.2	10.7	2.1	-6.9	-1.4	-2.1	-8.6
$\bar{X}_8$	6.6	0.7	3.9	1.2	-7.6	0.7	0.6	-6.9
$\bar{X}_9$	31.2	-0.5	20.2	4.4	-22.1	-0.6	-1.9	-24.8
relRMSE (×100)								
$\bar{X}_1$	7.6	2.8	6.0	3.8	7.3	2.7	2.8	6.6
$\bar{X}_2$	94.8	8.4	73.5	59.6	74.0	8.6	12.4	14.8
$\bar{X}_3$	6.7	3.6	4.4	3.7	11.1	3.6	3.7	6.0
$\bar{X}_4$	7.7	4.0	9.1	6.3	6.4	4.0	4.1	10.4
$\bar{X}_5$	96.2	2.7	17.9	4.5	6.0	2.8	3.1	8.4
$\bar{X}_6$	185.7	3.7	34.0	8.2	4.2	3.7	3.7	7.7
$\bar{X}_7$	65.6	3.2	12.7	3.9	7.6	3.3	3.6	9.1
$\bar{X}_8$	8.2	3.1	5.7	3.4	8.4	3.1	3.2	7.9
$\bar{X}_9$	36.3	4.8	23.9	7.4	22.6	4.9	5.4	25.1
95% CI Coverage								
$\bar{X}_1$	73.8	95.8	72.6	86.8	25.0	95.4	96.2	42.6
$\bar{X}_2$	15.4	95.4	23.4	34.6	7.2	94.2	76.0	84.6
$\bar{X}_3$	96.4	96.2	95.6	93.8	11.4	95.8	97.0	70.0
$\bar{X}_4$	87.0	94.8	78.4	87.2	67.2	95.4	94.6	27.4
$\bar{X}_5$	2.0	92.4	61.8	86.2	32.2	91.4	85.4	8.2
$\bar{X}_6$	23.4	93.0	78.2	87.4	86.0	93.6	91.0	40.8
$\bar{X}_7$	42.2	92.2	83.8	94.0	24.4	91.2	87.6	11.0
$\bar{X}_8$	84.8	93.8	90.4	93.6	20.6	94.8	94.8	30.0
$\bar{X}_9$	56.8	95.4	51.2	85.2	0.8	94.4	90.6	0.0

Table G.2: Summaries of the estimators of regression coefficients across 500 simulations in the outlier detection simulation. The first and second columns display results of the editing methods without an outlier detection step, which are exactly the same as the sixth and fourth columns of Table 2 of the main text, respectively. The value of C indicates the threshold parameter of the outlier detection method.

			With Outlier Detection					
			BE-min			BE		
	BE-min	BE	C=6	C=4	C=2	C=6	C=4	C=2
relBias (×100)								
β_0	2.8	0.9	5.4	7.6	9.2	0.7	-1.9	-8.8
β_1	39.7	-2.9	24.7	20.7	55.6	-2.9	-3.4	36.2
β_2	-24.7	1.7	-3.9	-2.7	-21.0	1.8	4.0	-1.7
β_3	-9.0	-0.4	-16.3	-16.9	-26.6	-0.3	0.9	-12.1
relRMSE (×100)								
β_0	12.3	9.4	11.8	12.2	15.8	9.2	10.2	17.6
β_1	43.2	16.3	29.8	26.0	57.4	16.5	16.7	40.4
β_2	28.0	11.9	13.9	12.6	25.6	11.9	13.3	17.4
β_3	15.6	11.0	20.2	20.0	29.1	10.9	11.2	18.8
95% CI Coverage								
β_0	89.4	95.2	88.0	86.8	83.6	96.6	95.0	85.0
β_1	34.6	94.0	66.8	72.0	3.0	94.2	92.4	45.4
β_2	37.8	93.2	90.0	93.8	67.0	93.2	92.4	92.6
β_3	83.0	94.2	68.2	63.2	33.4	94.0	94.8	80.0

Table H.1: Variables used in the CM edit rules. The upper table shows the ranges of reported values $(\tilde{L}_j, \tilde{U}_j)$ we introduced and whether a variable is subject to ratio edits. The values of the ratio edits provided by the Census Bureau are not displayed due to confidential reasons. The lower table displays the six balance edits.

Var. Name	TC	TIB	TIE	TVS	PW	OE	TE	WW	OW
$\tilde{L}_j$	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
$\tilde{U}_j$	10^6	10^5	10^5	10^6	10000	10000	10000	10^6	10^5
Ratio edit?	Y	Y	Y	Y	Y	Y	Y	Y	Y

Var. Name	SW	BEN	PH	C_a	C_b	C_c	C_d	C_e	IB_a
$\tilde{L}_j$	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
$\tilde{U}_j$	10^6	10^5	10000	10^6	10^5	10000	10000	10^5	10^5
Ratio edit?	Y	Y	Y	N	N	N	N	N	N

Var. Name	IB_b	IB_c	IE_a	IE_b	IE_c	PW_1	PW_2	PW_3	PW_4
$\tilde{L}_j$	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
$\tilde{U}_j$	10^5	10^5	10^5	10^5	10^5	10000	10000	10000	10000
Ratio edit?	N	N	N	N	N	N	N	N	N

Balance Edits

$TC = C_a + C_b + C_c + C_d + C_e$	$TIB = IB_a + IB_b + IB_c$
$TIE = IE_a + IE_b + IE_c$	$PW = (PW_1 + PW_2 + PW_3 + PW_4)/4$
$TE = PW + OE$	$SW = WW + OW$

REFERENCES

Banff Support Team (2007), "Functional Description of the Banff System for Edit and Imputation, Version 2.02 ," Technical Report, Statistics Canada.

Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H., and Karr, A. F. (2014), "Multiple Imputation of Missing or Faulty Values Under Linear Constraints," *Journal of Business & Economic Statistics*, 32, 375–386.

Kovar, J., Whitridge, P., and MacMillan, J. (1988), "Generalized Edit and Impuation System for Economic Surveys at Statistics Canada," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 627–630.

Kozak, R. (2005), "The Banff System for Automated Editing and Imputation," in *Proceedings of the Survey Methods Section*, pp. 1–10.

Woo, M. J., Reiter, J. P., Oganian, A., and Karr, A. F. (2009), "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation," *Journal of Privacy and Confidentiality*, 1, 111–124.

Supplementary Materials for "Simult. Edit-Imputation for Cont. Microdata"

Notations

$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$	true values of p variables for data subject i
$\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})$	reported values corresponding to $\boldsymbol{x}_i$
$\boldsymbol{s}_i = (s_{i1}, \ldots, s_{ip})$	error indicator corresponding to $\boldsymbol{y}_i$
(L_j, U_j)	range restriction for variable j
$(L_{j,j'}, U_{j,j'})$	ratio edit for variables j and j'
B_l	l-th balance edit, $l = 1, \ldots, q$
$\mathcal{B}_{\text{pass}}$	observed set of passed balanced edits
T_l	total variable involved in l-th balance edit. x_{iT_l} and y_{iT_l} are true and reported values of T_l of subject i
B_l	index of component variables involved in l-th balance edit. Note that for notational simplicity we use B_l to represent both the l-th balance edit and the indexes of its component variables.
NT	set of all-but-total variables, i.e., $\{1, \ldots, p\} \setminus \{T_l : l = 1, \ldots, q\}$ $\boldsymbol{x}_{i,NT}$ and $\boldsymbol{y}_{i,NT}$ are true and reported values of variables in NT
$\mathcal{X}$	region of all potential records that passes all edits
$\mathcal{D}$	region of all potential records that passes all inequality constraints. Note that $\mathcal{X} = \mathcal{D} \cap \{\boldsymbol{x} : \sum_{j \in B_l} x_j = x_{T_l}, \forall l\}$
$\mathcal{Y}$	support of reported values $\boldsymbol{y}_i$. Note that generally $\mathcal{X} \subset \mathcal{Y}$ for editing of continuous data
$(\tilde{L}_j, \tilde{U}_j)$	range of reported values of variable j. Note that $(\tilde{L}_j, \tilde{U}_j)$ may or not equal to (L_j, U_j)
A_i	indicator showing how record i violates edit constraints. In our model, $A_i = 0$ if i satisfies all edits $A_i = 1$ if i fails at least one balance edit and $\boldsymbol{x} \in \mathcal{D}$ $A_i = 2$ if i passes all balance edits but $\boldsymbol{x} \notin \mathcal{D}$ $A_i = 3$ if i fails at least one balance edit and $\boldsymbol{x} \notin \mathcal{D}$.

$\boldsymbol{\theta}$	parameter of true data model $f\left(\boldsymbol{x}_{i,NT}\vert\boldsymbol{\theta}\right)$
ψ_s	parameter of error indicator model $f(\boldsymbol{s}_i, A_i \mid \boldsymbol{x}_i, \psi_s)$
ψ_y	parameter of reporting model $f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{s}_i, A_i, \psi_y)$. Note that $\dot{f}(\boldsymbol{y}_i\vert\boldsymbol{x}_i, \boldsymbol{s}_i, A_i, \psi_y)$ is an unnormalized density of $f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{s}_i, A_i, \psi_y)$
$f(\boldsymbol{x}_{i,NT}\vert\boldsymbol{\theta})$	joint density of true values for all-but-total variables
$\boldsymbol{y}_i^{UF}$	correctly reported (not flagged) values given $(\boldsymbol{x}_i, \boldsymbol{s}_i, A_i)$, i.e., $\{y_{ij} : s_{ij} = 0, j = 1, \ldots, p\}$
$\boldsymbol{y}_i^{F}$	incorrectly reported (flagged) values given $(\boldsymbol{x}_i, \boldsymbol{s}_i, A_i)$, i.e., $\{y_{ij} : s_{ij} = 1, j = 1, \ldots, p\}$
$\boldsymbol{x}_i^{UF}, \boldsymbol{x}_i^{F}$	true values corresponding to $\boldsymbol{y}_i^{UF}$ and $\boldsymbol{y}_i^{F}$
$\mathcal{C}_i$	set of variables known to have $s_{ij} = 0$ with corresponding values $\boldsymbol{y}_i^{\mathcal{C}_i} = \boldsymbol{x}_i^{\mathcal{C}_i}$ and $\boldsymbol{s}_i^{\mathcal{C}_i} = 0$.
$\mathcal{E}_i$	set of remaining values for record i, i.e., all cases with missing or possibly erroneous y_{ij}, with corresponding values $(\boldsymbol{x}_i^{\mathcal{E}_i}, \boldsymbol{s}_i^{\mathcal{E}_i}, \boldsymbol{y}_i^{\mathcal{E}_i})$.
$\boldsymbol{X}_n, \boldsymbol{S}_n, \boldsymbol{Y}_n$	collection of values for n subjects: $\boldsymbol{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{S}_n = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$ and $\boldsymbol{Y}_n = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$
$\mathcal{S}(\boldsymbol{y}_i, A_i)$	set of feasible $\boldsymbol{s}_i$ for subject i, which is function of $\boldsymbol{y}_i$ and A_i
$\boldsymbol{X}_{N_{\text{aug}}-n}$	set of $\boldsymbol{x}_i$ for $N_{\text{aug}} - n$ hypothetical, unobserved individuals, which are used to estimate parameters of $\boldsymbol{X}_n$ from an unconstrained distribution of $\boldsymbol{X}_{\text{aug}} = (\boldsymbol{X}_n, \boldsymbol{X}_{N_{\text{aug}}-n})$
z_i	mixture membership indicator of record i, i.e., $z_i \in \{1, \ldots, K\}$ when the mixture of K distributions is used
π_k	parameters for z_i, i.e., $Pr(z_i = k) = \pi_k$ where $\sum_k \pi_k = 1$
$\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$	mean vector and covariance matrix of normal distribution for $\log(\boldsymbol{x}_{i,NT})$ in the k-th mixture component
$\mu, \Sigma, \boldsymbol{\pi}$	parameters of a mixture of distributions defined as $\mu = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$, $\Sigma = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$

$\boldsymbol{\mu}_0, h_0$	parameters for $\boldsymbol{\mu}_k$
$\boldsymbol{\Phi}, \zeta_0, a_\Phi, b_\Phi$	parameters for $\boldsymbol{\Sigma}_k$ where $\boldsymbol{\Phi} = diag(\Phi_1, \ldots, \Phi_{p-q})$
$\alpha, a_\alpha, b_\alpha$	parameters for π_k (or ν_k)
$\boldsymbol{X}^{\text{pop}}$	population used in the simulation study of Section 4 comprising 1000000 records measured on $p = 9$ variables
$\boldsymbol{X}^r$	r-th random sample of size $n = 1000$ drawn from $\boldsymbol{X}^{\text{pop}}$, $r = 1, \ldots, R$
$\boldsymbol{Y}^r$	reported dataset corresponding to $\boldsymbol{X}^r$
$\boldsymbol{X}^{r(m)}$	m-th corrected (completed) dataset corresponding to $\boldsymbol{Y}^r$ (and $\boldsymbol{X}^r$), $m = 1, \ldots, M$
$\boldsymbol{X}^{r,\text{pass}}$	sample of edit-passing records only, corresponding to $\boldsymbol{Y}^r$
ω	pre-determined small number replacing nonpositive values, to work with ratio edits. In our example, $\omega = 0.1$
w_j	reliability weight of variable j
$\delta(\cdot)$	Dirac delta function with the point mass at zero
$I[\cdot]$	indicator with the value one if the statement inside the brackets is true and zero otherwise
Bernoulli(p)	Bernoulli distribution with success probability p
Beta(a, b)	Beta distribution with mean $a/(a+b)$
Binomial(n, p)	Binomial distribution with mean np
Categorical$(p_1, \ldots, p_K)$	Categorical distribution with the event probabilities $p_1, \ldots, p_K$ where $\sum_k p_k = 1$
Gamma(a, b)	Gamma distribution with mean a/b
InverseWishart$(a, \boldsymbol{B})$	Inverse Wishart distribution with d.f. a and the scale matrix $\boldsymbol{B}$
N$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Normal distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. N$(\boldsymbol{x}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ indicates the density of N$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ evaluated at $\boldsymbol{x}$
Unif$(\mathcal{A})$	Uniform distribution on the support $\mathcal{A}$

Abbreviations

AAI	all active items
CM	Census of Manufactures
FAR	faulty at random
F-H	Fellegi and Holt
MFI	minimum fields to impute
MWFI	minimum weighted fields to impute
NFAR	not faulty at random
RMSE	root mean squared errors