Online Supplement to Bayesian Simultaneous Edit and Imputation for Multivariate Categorical Data

Daniel Manrique-Vallier and Jerome P. Reiter*

August 3, 2016

This supplement includes the algorithm for processing table slice definitions used in the MCMC sampler for edit imputation with categorical data; see Section 1. It also includes results from additional simulations referred to in the main text (Section 2), and the list of slice definitions that describe the edits (Section 3).

1 Algorithm for processing slice definitions

Here we describe a slightly modified version of the algorithm developed in Manrique-Vallier and Reiter (2014) for transforming a collection of slice definitions into a collection of disjoint slice definitions whose union represent the same set of cells. Let $\Omega \subset \mathcal{C}^*$ be a collection of potentially overlapping slice definitions. We seek an alternative collection $\Gamma \subset \mathcal{C}^*$ such that

$$\bigcup_{\boldsymbol{\omega}\in\Omega} \overline{\boldsymbol{\omega}} = \bigcup_{\boldsymbol{\mu}\in\Gamma} \overline{\boldsymbol{\mu}}$$

and $\overline{\boldsymbol{\mu}_1} \cap \overline{\boldsymbol{\mu}_2} = \emptyset$ for every $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \Gamma$. See the main text for definitions of $\boldsymbol{\mu}$ and $\overline{\boldsymbol{\mu}}$.

The algorithm from Manrique-Vallier and Reiter (2014) is akin to the Gram-Schmidt orthogonalization procedure. It proceeds by a repeated application of a formal set-subtraction operation in the space of slice definitions, exploiting the fact that for any two sets A and B the difference set $A \setminus B$ and B are disjoint, but $(A \setminus B) \cup B = A \cup B$. An example adapted from Manrique-Vallier and Reiter

*Daniel Manrique-Vallier is Assistant Professor of Statistics, Indiana University, Bloomington, IN 47408 (e-mail dmanriqu@indiana.edu); and Jerome P. Reiter is Professor of Statistical Science, Duke University, Durham, NC 27708-0251, (e-mail: jerry@stat.duke.edu). This research was supported by a grant from the National Science Foundation (SES-11-31897).

(2014) can help to illustrate the main ideas. Let $\boldsymbol{\mu}_1 = (1, 1, *, *)$ and $\boldsymbol{\mu}_2 = (1, *, 2, *)$ be two slice definitions, and assume that the number of levels in x_3 is $L_3 = 2$. The corresponding slices are not disjoint since $\overline{\boldsymbol{\mu}_1} \cap \overline{\boldsymbol{\mu}_2} = \overline{\text{int}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)} = \overline{(1, 1, 2, *)}$. However, $\overline{\boldsymbol{\mu}_1} = \overline{(1, 1, *, *)} = \overline{(1, 1, 1, *)} \cup \overline{(1, 1, 2, *)}$, so the collection $\{(1, 1, 1, *), (1, 1, 2, *)\}$ is *equivalent* to the table slice $\overline{\boldsymbol{\mu}_1}$. Moreover, since the intersection slice definition, $(1, 1, 2, *)$, appears explicitly in the new representation of $\overline{\boldsymbol{\mu}_1}$, we can get a representation of the set $\overline{\boldsymbol{\mu}_1} \setminus \overline{\boldsymbol{\mu}_2}$ in terms of slice definitions by simply removing intersection slice, resulting in the collection $\{(1, 1, 1, *)\}$. Finally, we can get a collection of disjoint slices equivalent to $\overline{\boldsymbol{\mu}_1} \cup \overline{\boldsymbol{\mu}_2}$ by merging the collection corresponding to $\overline{\boldsymbol{\mu}_1} \setminus \overline{\boldsymbol{\mu}_2}$ and $\{\boldsymbol{\mu}_1\}$, resulting in $\{(1, 1, 1, *), (1, 1, 2, *)\}$.

To implement this idea, we define the *expansion of slice definition* $\boldsymbol{\mu}$ *with respect to the set of coordinate indexes* $\xi \subset \{1, 2..., J\}$ as the collection of slice definitions, $Expan(\boldsymbol{\mu}, \xi) \subset \mathcal{C}^*$, such that

$$Expan(\boldsymbol{\mu}, \xi) = \{(x_1, ..., x_J) : x_j = \mu_j \text{ if } j \notin \xi \text{ or } \mu_j \neq *.\ x_j \in \{1, ..., L_j\} \text{ otherwise}\}.$$

This operation simply turns a slice definition into an equivalent collection of slice definitions by replacing the placeholder component * with actual discrete levels. For example, let $L_3 = 2$ and $L_5 = 3$. We have that

$$\begin{aligned} Expan((4, 1, *, *, *), \{3, 5\}) = \{&(4, 1, 1, *, 1), (4, 1, 1, *, 2), (4, 1, 1, *, 3), \\ &(4, 1, 2, *, 1), (4, 1, 2, *, 2), (4, 1, 2, *, 3)\}. \end{aligned}$$

Using this operation we can now describe the algorithm for transforming a collection of slice definitions Ω into an equivalent collection Γ with disjoint elements:

1. Sort the elements of Ω into a sequence $(\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_M)$ in decreasing order on the number of cells represented by each slice definition: $|\overline{\boldsymbol{\omega}_1}| \geq |\overline{\boldsymbol{\omega}_2}| \geq ... \geq |\overline{\boldsymbol{\omega}_M}|$.
2. Let $\Gamma \leftarrow \{\boldsymbol{\omega}_1\}$.
3. For $i = 2, ..., M$ do
   (a) Let $\Theta \leftarrow \{\boldsymbol{\omega}_k : k < i \text{ and } \overline{\text{int}(\boldsymbol{\omega}_i, \boldsymbol{\omega}_k)} \neq \emptyset\}$

(b) Let $\Xi \leftarrow \{j : \theta_j \neq * \text{ and } \mu_j = * \text{ for any } (\theta_1, ..., \theta_J) \in \Theta\}$

(c) If either Θ or Ξ are empty

- Let $\Gamma \leftarrow \Gamma \cup \{\boldsymbol{\omega}_i\}$

Else

- Let $\Gamma \leftarrow \Gamma \cup \{\text{All elements of } Expan(\boldsymbol{\omega}_i, \Xi) \text{ that are disjoint with every element of } \Theta\}$.

Finally, we note that step 3 (a) can be greatly sped up by using the fact that $\overline{\text{int}(\boldsymbol{\omega}_i, \boldsymbol{\omega}_k)} = \emptyset$ is equivalent to the statement, $\omega_{ij} \neq \omega_{kj}$ for some $j \in \{1, ..., J\}$ such that $\omega_{ij} \neq *$ and $\omega_{kj} \neq *$.

2 Results From Additional Empirical Studies

2.1 Experiment 1: Repeated sampling with uniform $\epsilon = 0.1$ contamination

In this experiment, described in Section 4.1 of the main text, we take 500 random samples with $n = 1000$ from the PUMS data and contaminate each using a reporting model with independent errors model with rate $\epsilon = 0.1$ and uniform substitution. We fit the EI-DPM with the weak prior $\epsilon \sim \text{Beta}(1, 1)$ to each of the samples, and estimate the 1,824 3-way estimable margins using 50 multiply edited-imputed datasets obtained from the posterior distribution. For comparison we also compute the same quantities from 50 multiply edited-imputed datasets using the F-H error localization method.

Figure 1 shows the average over the 500 trials of the multiple imputation estimates versus their actual population values, for the EI-DPM and the F-H methods. Figure 2 shows the empirical coverage of 95% intervals corresponding to each of the estimands, computed from the multiply edited-imputed datasets generated EI-DPM and the F-H approaches, versus the corresponding coverage obtained from multiply imputing of the faulty values knowing their true location (labeled "Oracle"). Both figures indicate that EI-DPM results in better performance than the F-H procedure at this error rate.

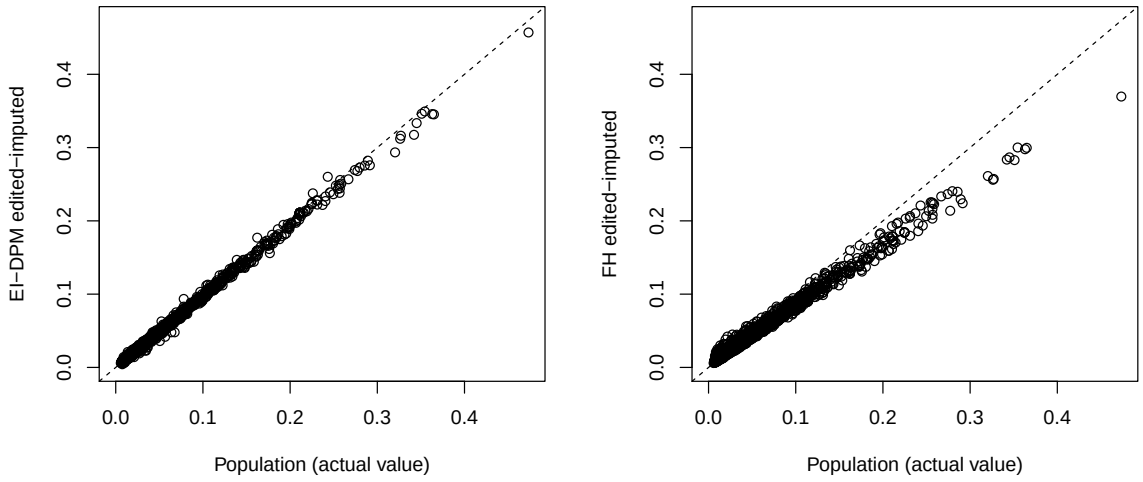

Figure 1: **Experiment 1:** Average over 500 replications of the multiple imputation estimates of 3-way margin proportions versus their actual population values, for faulty data with $\epsilon = 0.1$. EI-DPM in left panel. F-H in right panel.

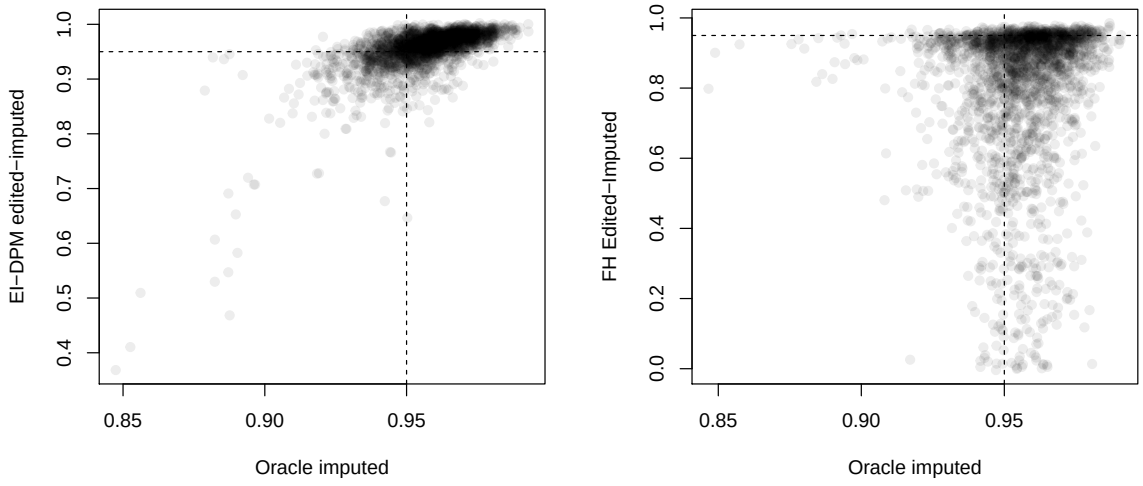

Figure 2: **Experiment 1:** Empirical coverage over 500 replications of multiple-imputation 95% intervals for 1,824 3-way margin proportions obtained from contaminated samples with $\epsilon = 0.1$. Left panel: EI-DPM edited-imputed vs "Oracle" imputed samples. Right panel: Fellegi-Holt vs "Oracle" imputed. Discontinuous lines mark nominal 95% coverage levels. Unif(-0.005, 0.005) noise added for clarity.

Prior	Probability
$\epsilon \sim \text{Beta}(1, 10^4)$	36.2%
$\epsilon \sim \text{Beta}(1, 10^5)$	5.59%
$\epsilon \sim \text{Beta}(1, 10^6)$	0.578%

Table 1: **Experiment 3:** Probability of detecting at least one error in contaminated records without detectable errors for a single contaminated dataset with $n = 1000$ and $\epsilon = 0.4$. The probability of detecting errors in records that violate edit rules is 100%.

Method	Mean absolute relative error
EI-DPM $\epsilon \sim \text{Beta}(1, 10^4)$	0.322
EI-DPM $\epsilon \sim \text{Beta}(1, 10^5)$	0.365
EI-DPM $\epsilon \sim \text{Beta}(1, 10^6)$	0.375
Fellegi-Holt	0.393
Raw data	0.404
Uncontaminated (*)	0.143

Table 2: **Experiment 2:** Mean absolute relative error of estimates of 1843 3-way margin proportions obtained from a single contaminated dataset with $\epsilon = 0.4$ using five different methods: EI-DPM with three different strong priors, Fellegi-Holt, and the raw contaminated data. We also include the estimates obtained from the dataset before contamination (*).

2.2 Experiment 2: Single trial with $\epsilon = 0.4$ and strong priors for ϵ

In this experiment, reported in Section 4.2 of the main text, we apply the edit-imputation methods on a single contaminated sample of size $n = 1000$ using different strong priors on ϵ. The contamination follows the independent error localization with $\epsilon = 0.4$, and uniform substitution model. We investigate the impacts of assuming $\epsilon \sim \text{Beta}(1, 10^4)$, $\epsilon \sim \text{Beta}(1, 10^5)$, and $\epsilon \sim \text{Beta}(1, 10^6)$.

Table 1 shows the rate of detection of errors in records that do not contain violations to the edit rules for the EI-DPM with different priors. As expected, the higher the value of b_ϵ (suggesting that records do not contain errors), the lower the error detection rate. Lower error detection implies less editing, effectively reducing the number of changes to records with no obvious errors. Thus, we can use the prior specification of ϵ for modulating the desired aggressiveness of the EI-DPM procedure.

Table 2 shows the mean absolute relative error of all estimands, $\sum_{i=1}^{1824} |(\hat{\theta}_i - \theta_i)/\theta_i|/1824$, where θ_i is the actual population value of i-th estimand and $\hat{\theta}_i$ is its corresponding estimate. As expected, the stronger the prior suggests that records do not contain errors, the worse the estimates. However,

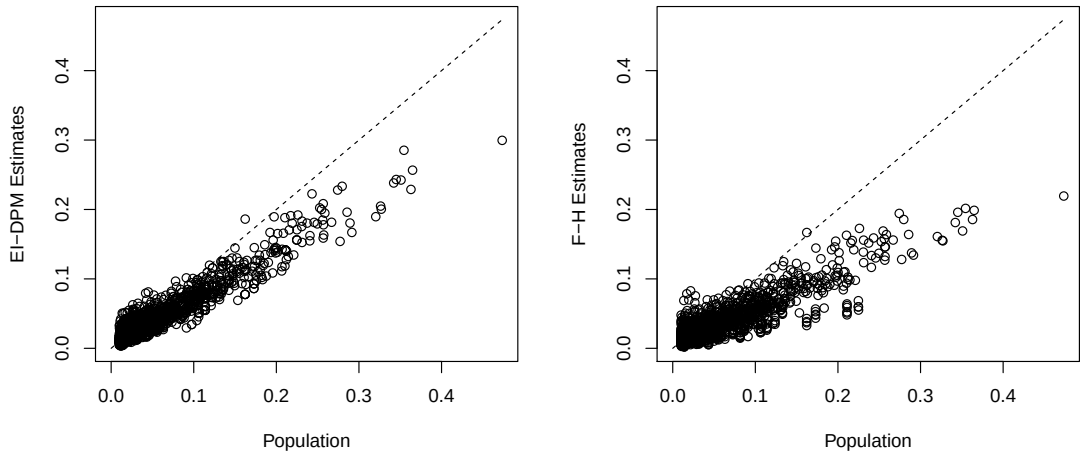

Figure 3: **Experiment 2:** Multiple edit-imputation estimates of 3-way margin proportions versus their actual population values for a single contaminated sample with $\epsilon = 0.4$. The left corresponds to the EI-DPM method with a strong $\epsilon \sim \text{Beta}(1, 10^4)$ prior; the right panel to the Fellegi-Holt method.

even with the extremely conservative prior specification $\epsilon \sim \text{Beta}(1, 10^6)$, the EI-DPM estimates tend to have smaller percentage errors than those based on the F-H approach. Figures 3, 4 and 5 show the estimates versus their actual values for the EI-DPM with the three strong priors and for the F-H method.

2.3 Experiment 3: Repeated sampling with uniform $\epsilon = 0.4$ contamination and no undetectable errors

In this experiment, reported in Section 4.2 of the main text, we take 500 random samples of size $n = 1000$ from the PUMS data and contaminate each using a reporting model with independent errors with rate $\epsilon = 0.4$ and uniform substitution. Then we reset all the records without detectable errors to their original values, so that the resulting datasets do not include undetectable errors. We fit the EI-DPM models with $\epsilon \sim \text{Beta}(1, 1)$, after setting the error indicators of records that do not violate edits (i.e., no detectable errors) to $\mathbf{E}_i = (0, ..., 0)$ a priori. We compute the 1824 3-way margin proportions using 50 multiply edited-imputed datasets obtained from the posterior

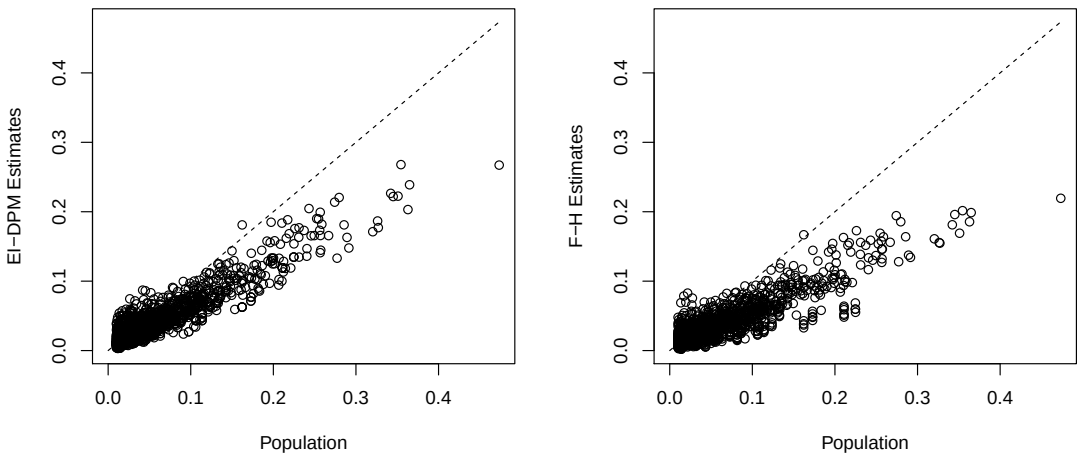

Figure 4: **Experiment 3:** Multiple edit-imputation estimates of 3-way margin proportions versus their actual population values for a single contaminated sample with $\epsilon = 0.4$. The left corresponds to the EI-DPM method with a strong $\epsilon \sim \text{Beta}(1, 10^5)$ prior; the right panel to the Fellegi-Holt method.

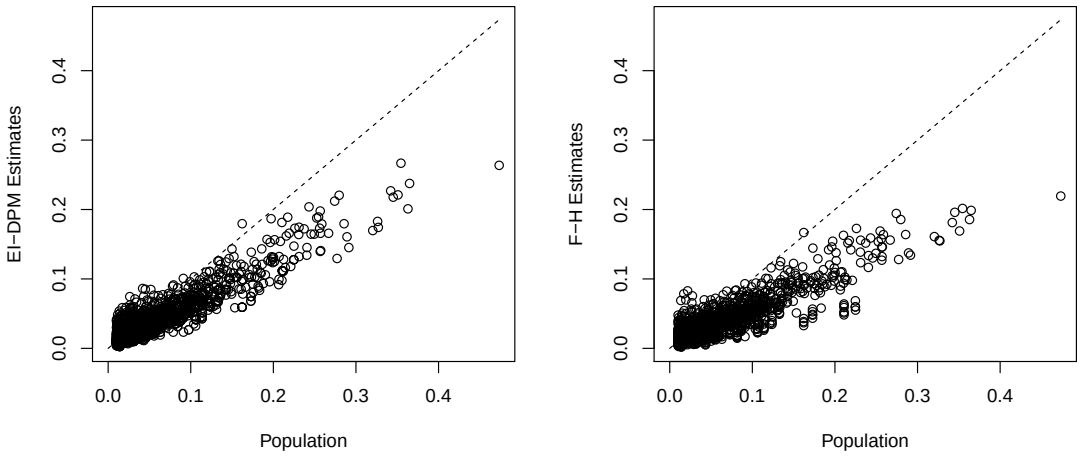

Figure 5: **Experiment 3:** Multiple edit-imputation estimates of 3-way margin proportions versus their actual population values for a single contaminated sample with $\epsilon = 0.4$. The left corresponds to the EI-DPM method with a strong $\epsilon \sim \text{Beta}(1, 10^6)$ prior; the right panel to the Fellegi-Holt method.

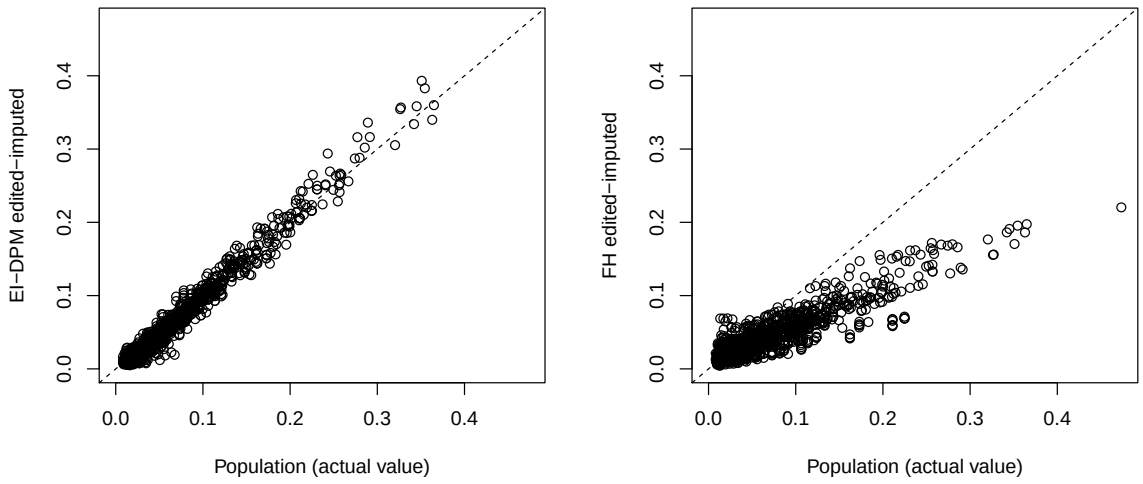

Figure 6: **Experiment 3:** Average over 500 replications of the multiple imputation estimates of 3-way margin proportions versus their actual population values, for faulty data with $\epsilon = 0.4$ and no undetectable errors. EI-DPM in left panel. F-H in right panel.

distribution. For comparison we also compute the same quantities from 50 multiply edited-imputed datasets using the F-H error localization method.

Figure 6 shows the average over the 500 trials of the multiple-imputation estimates versus their actual population values, for the EI-DPM and the F-H methods. Figure 7 shows the empirical coverage of 95% intervals corresponding to each of the estimands, computed from the multiply edited-imputed datasets generated EI-DPM and the F-H approaches, versus the corresponding coverage obtained from multiply imputing of the faulty values at their true locations (labeled "Oracle"). The EI-DPM approach results in better performance than the EI-DPM approach in this scenario.

2.4 Experiment 4 : Repeated sampling with misspecified model and no undetectable errors

In this experiment, reported in Section 4.2 of the main text, we take 500 random samples of size $n = 1000$ from the PUMS data and contaminate each using a reporting model with different per-

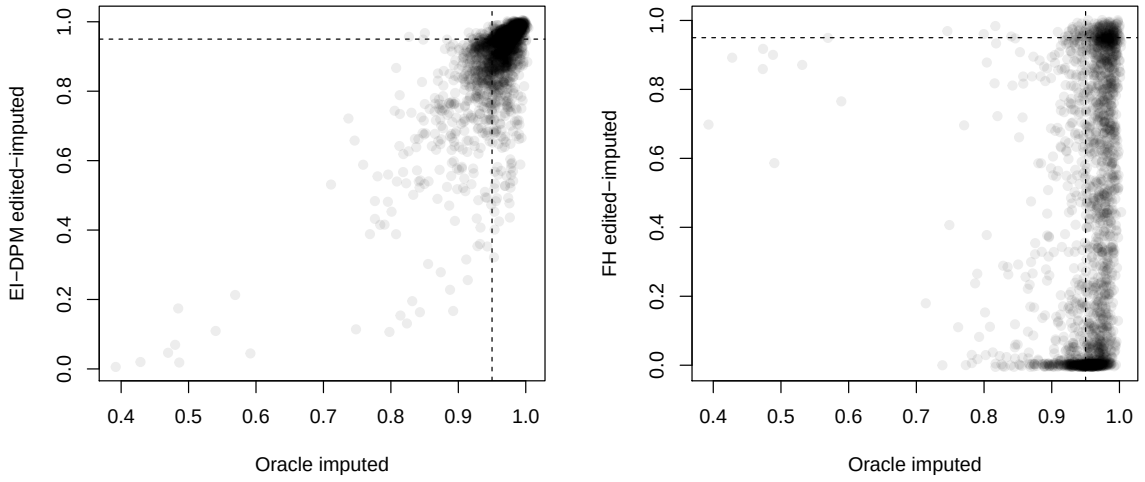

Figure 7: **Experiment 3:** Empirical coverage over 500 replications of multiple-imputation 95% intervals for 1,824 3-way margin proportions obtained from contaminated samples with $\epsilon = 0.4$ and no undetectable errors. Left panel: EI-DPM edited-imputed vs "Oracle" imputed samples. Right panel: Fellegi-Holt vs "Oracle" imputed. Discontinuous lines mark nominal 95% coverage levels. Unif(-0.005, 0.005) noise added for clarity.

variable rates $\epsilon \overset{iid}{\sim} \text{Beta}(5, 50)$, and non-uniform substitution model with probabilities proportional to samples from a Gamma$(40, 1)$ distribution. Then we reset all the records without detectable errors to their original values, so that the resulting datasets do not include undetectable errors. We fit misspecified EI-DPM models with common error rate ϵ with prior distribution $\epsilon \sim \text{Beta}(1, 1)$. Additionally, we prevent modification of records with non-detectable errors by setting the error indicators of records that do not violate edits to $\mathbf{E}_i = (0, ..., 0)$ a priori. We compute the 1824 3-way margin proportions using 50 multiply edited-imputed datasets obtained from the posterior distribution. For comparison we also compute the same quantities from 50 multiply edited-imputed datasets using the F-H error localization method.

Figure 8 shows the empirical coverage of 95% intervals corresponding to each of the estimands, computed from the multiply edited-imputed datasets generated EI-DPM and the F-H approaches, versus the corresponding coverage obtained from multiply imputing of the faulty values at their true locations (labeled "Oracle"). A plot comparing the estimates themselves can be found in the main text.

3 Edits Definition

Table 3 presents the complete list of 60 table slice definitions that describe the edits for the extract of the 2000 U.S. census data for the state of New York used in this article. The slices described by these definitions are not disjoint. For example the first slice, defined by $\boldsymbol{\mu}_1 = (1, 0, *, *, *, *, *, *, *, *)$, and slice defined by $\boldsymbol{\mu}_8 = (*, *, < 15, *, 1, *, *, *, *, *)$ intersect in the set defined by $(1, 0, < 15, *, 1, *, *, *, *, *)$. After applying the algorithm from Section 1 of this document, this collection gets transformed into an equivalent collection of 567 disjoint slice definitions. Table 4 details the variables and codes used in this dataset.

	OWNERSHP	MORTGAGE	AGE	SEX	MARST	RACESING	EDUC	EMPSTAT	DISABWRK	VETSTAT
1	1	0	*	*	*	*	*	*	*	*
2	0	1	*	*	*	*	*	*	*	*
3	2	1	*	*	*	*	*	*	*	*
4	0	3	*	*	*	*	*	*	*	*
5	2	3	*	*	*	*	*	*	*	*
6	0	4	*	*	*	*	*	*	*	*
7	2	4	*	*	*	*	*	*	*	*
8	*	*	<15	*	1	*	*	*	*	*
9	*	*	<15	*	2	*	*	*	*	*
10	*	*	<15	*	3	*	*	*	*	*
11	*	*	<15	*	4	*	*	*	*	*
12	*	*	<15	*	5	*	*	*	*	*
13	*	*	[18,24]	*	*	*	*	0	*	*
14	*	*	[25,35]	*	*	*	*	0	*	*
15	*	*	[36,50]	*	*	*	*	0	*	*
16	*	*	[51,70]	*	*	*	*	0	*	*
17	*	*	>70	*	*	*	*	0	*	*
18	*	*	16	*	*	*	*	0	*	*
19	*	*	17	*	*	*	*	0	*	*
20	*	*	<15	*	*	*	*	1	*	*
21	*	*	15	*	*	*	*	1	*	*
22	*	*	<15	*	*	*	*	2	*	*
23	*	*	15	*	*	*	*	2	*	*
24	*	*	<15	*	*	*	*	3	*	*
25	*	*	15	*	*	*	*	3	*	*
26	*	*	<15	*	*	*	7	*	*	*
27	*	*	15	*	*	*	7	*	*	*
28	*	*	<15	*	*	*	8	*	*	*
29	*	*	15	*	*	*	8	*	*	*
30	*	*	<15	*	*	*	10	*	*	*
31	*	*	15	*	*	*	10	*	*	*
32	*	*	16	*	*	*	10	*	*	*
33	*	*	17	*	*	*	10	*	*	*
34	*	*	<15	*	*	*	11	*	*	*
35	*	*	15	*	*	*	11	*	*	*
36	*	*	16	*	*	*	11	*	*	*
37	*	*	17	*	*	*	11	*	*	*
38	*	*	[18,24]	*	*	*	*	*	0	*
39	*	*	[25,35]	*	*	*	*	*	0	*
40	*	*	[36,50]	*	*	*	*	*	0	*
41	*	*	[51,70]	*	*	*	*	*	0	*
42	*	*	>70	*	*	*	*	*	0	*
43	*	*	16	*	*	*	*	*	0	*
44	*	*	17	*	*	*	*	*	0	*
45	*	*	<15	*	*	*	*	*	1	*
46	*	*	15	*	*	*	*	*	1	*
47	*	*	<15	*	*	*	*	*	4	*
48	*	*	15	*	*	*	*	*	4	*
49	*	*	[18,24]	*	*	*	*	*	*	0
50	*	*	[25,35]	*	*	*	*	*	*	0
51	*	*	[36,50]	*	*	*	*	*	*	0
52	*	*	[51,70]	*	*	*	*	*	*	0
53	*	*	>70	*	*	*	*	*	*	0
54	*	*	17	*	*	*	*	*	*	0
55	*	*	<15	*	*	*	*	*	*	1
56	*	*	15	*	*	*	*	*	*	1
57	*	*	16	*	*	*	*	*	*	1
58	*	*	<15	*	*	*	*	*	*	2
59	*	*	15	*	*	*	*	*	*	2
60	*	*	16	*	*	*	*	*	*	2

Table 3: Complete list of slice definitions for the NY data before processing.

Variable	Code	Description
OWNERSHP	0	N/A
	1	Owned or being bought (loan)
	2	Rented
MORTGAGE	0	N/A
	1	No, owned free and clear
	3	Yes, mortgaged/ deed of trust or similar debt
	4	Yes, contract to purchase
SEX	1	Male
	2	Female
AGE	<15	Less than 15
	15	Fifteen
	16	Sixteen
	17	Seventeen
	[18,24]	Between 18 and 24
	[25,35]	Between 25 and 35
	[36,50]	Between 36 and 50
	[51,70]	Between 51 and 70
	>70	More than 70
MARST	1	Married, spouse present
	2	Married, spouse absent
	3	Separated
	4	Divorced
	5	Widowed
	6	Never married/single
RACESING	1	White
	2	Black
	3	American Indian/Alaska Native
	4	Asian and/or Pacific Islander
	5	Other race, non-Hispanic
EDUC	0	N/A or no schooling
	1	Nursery school to grade 4
	2	Grade 5, 6, 7, or 8
	3	Grade 9
	4	Grade 10
	5	Grade 11
	6	Grade 12
	7	1 year of college
	8	2 years of college
	10	4 years of college
	11	5+ years of college
EMPSTAT	0	N/A
	1	Employed
	2	Unemployed
	3	Not in labor force
DISABWRK	0	N/A
	1	No disability that affects work
	4	Difficulty working
VETSTAT	0	N/A
	1	Not a veteran
	2	Veteran

Table 4: Codes for levels of variables in the NY Data.

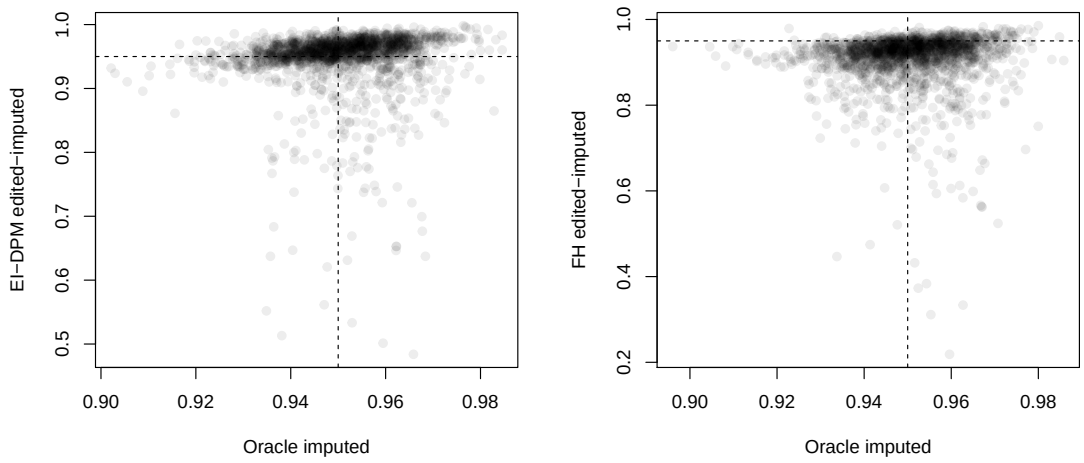

Figure 8: **Experiment 4:** Empirical coverage over 500 replications of multiple-imputation 95% intervals for 1,824 3-way margin proportions obtained from non-uniformly contaminated data without undetectable errors. Left panel: *Misspecified* EI-DPM edited-imputed vs "Oracle" imputed samples. Right panel: F-H vs "Oracle" imputed. Discontinuous lines mark nominal 95% coverage levels. Unif(-0.005, 0.005) noise added for clarity

References

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.