8 Appendix

8.1 Bayes Estimate

We calculate a Bayes estimate $\hat{\boldsymbol{Z}}$ for the linkage parameter $\boldsymbol{Z}$ by assigning different positive losses to different types of errors, and minimizing posterior expected loss. We adopt the loss function proposed in Sadinle (2017) in which $\hat{Z}_j \in \{1, \dots, n_1, n_1 + j, R\}$, with R representing the option to leave the matching undetermined by the model. Specifically, we have

$$L(\hat{Z}_j, Z_j) = \begin{cases} 0, & \text{if } Z_j = \hat{Z}_j; \\ \theta_R, & \text{if } \hat{Z}_j = R; \\ \theta_{10}, & \text{if } Z_j \leq 1, \hat{Z}_j = n_1 + j; \\ \theta_{01}, & \text{if } Z_j = n_1 + j, \hat{Z}_j \leq n_1; \\ \theta_{11}, & \text{if } Z_j \leq n_1, \hat{Z}_j \leq n_1, Z_j \neq \hat{Z}_j. \end{cases}$$

Here, θ_R is the loss from not making a decision on the linkage status, θ_{10} is the loss from a false non-match, θ_{01} is the loss from a false match, and θ_{11} is the loss from the special case of a false match in which the record has a true match other than the one estimated by the model.

In general, we set $(\theta_{10}, \theta_{01}, \theta_{11}, \theta_R) = (1, 1, 2, \infty)$ inducing the decision rule

$$\hat{Z}_j = \begin{cases} i, & \text{if } P(Z_j = i|\Gamma) > \frac{1}{2}; \\ 0, & \text{otherwise}. \end{cases}$$

Since `fabl` does not strictly enforce one-to-one matching, it is possible for this Bayes estimate to link multiple records in $\boldsymbol{X_2}$ to one record in $\boldsymbol{X_1}$. In the event that we have two records j and j' such that both $P(\hat{Z}_j = i|\Gamma) > \frac{1}{2}$ and $P(\hat{Z}_{j'} = i|\Gamma) > \frac{1}{2}$, we accept the match with the higher posterior probability, and declare the other to have no match. Since each Z_j is independent, this is equivalent to minimizing the expected loss subject to the constraint that $\hat{Z}_j \neq \hat{Z}_{j'}$ for all $j \neq j'$. A similar approach appears in the most probable maximal matching sets used by Steorts et al. (2016) to match records to latent entities.

When we seek a partial estimate of the linkage structure, leaving a portion of record pairs to be classified manually in clerical review, we adopt losses $(\theta_{10}, \theta_{01}, \theta_{11}, \theta_R) = (1, 1, 2, .1)$. For a more in-depth explanation of this function and the induced Bayes estimate, see Sadinle (2017).

8.2 Derivations of Full Conditionals

We provide detailed derivations of the full-conditionals provided in Section 3.1. The $\boldsymbol{m}$ and $\boldsymbol{u}$ parameters are updated through standard multinomial-Dirichlet distributions. For a particular m_{fl}, we have

$$p(m_{fl}|\Gamma, \mathbf{Z}) \propto \prod_{i=1}^{n_1} \prod_{j=1}^{n_2} m_{fl}^{I(Z_j=i)I(\gamma_{ij}^f=l)I_{obs}(\gamma_{ij}^f)} \times m_{fl}^{\alpha_{fl}-1}$$

$$= m_{fl}^{\alpha_{fl}(\boldsymbol{Z})-1},$$

where $\alpha_{fl}(\boldsymbol{Z}) = \alpha_{fl} + \sum_{i,j} I_{obs}(\gamma_{ij}^f) I(\gamma_{ij}^f = l) I(z_j = i)$. Analogous procedures lead to the posterior distribution $p(u_{fl}|\Gamma, \mathbf{Z}) \propto u_{fl}^{\beta_{fl}(\boldsymbol{Z})-1}$, where $\beta_{fl}(\boldsymbol{Z}) = \beta_{fl} + \sum_{i,j} I_{obs}(\gamma_{ij}^f) I(\gamma_{ij}^f = l) I(z_j \neq i)$. Thus for the vectors of parameters $\boldsymbol{m}_f$ and $\boldsymbol{u}_f$, we have

$$\begin{aligned}\boldsymbol{m}_f|\boldsymbol{Z}, \Gamma &\sim \text{Dirichlet}(\alpha_{f1}(\boldsymbol{Z}), \dots, \alpha_{fL_f}(\boldsymbol{Z})),\\ \boldsymbol{u}_f|\boldsymbol{Z}, \Gamma &\sim \text{Dirichlet}(\beta_{f1}(\boldsymbol{Z}), \dots, \beta_{fL_f}(\boldsymbol{Z})).\end{aligned}$$

In order to define the full conditional for $\boldsymbol{Z}$, we must provide the posterior distribution for π, and provide a more tractable expression for the portion of the likelihood relevant to a particular record $j \in \boldsymbol{X}_2$. Since π encodes the rate of matching across the two data files, the posterior distribution $p(\pi|\Gamma, \boldsymbol{Z}, \boldsymbol{m}, \boldsymbol{u}, \alpha_\pi, \beta_\pi)$ depends only on the number of links $n_{12}(\boldsymbol{Z}) = \sum_{i=1}^{n_2} I(z_j < n_1 + j)$ encoded by $\boldsymbol{Z}$ (and hyperparameters). Thus, we use $p(\pi|\mathbf{Z}, \alpha_\pi, \beta_\pi)$ and have

$$\begin{aligned}p(\pi|\boldsymbol{Z}, \alpha_\pi, \beta_\pi) &\propto p(\boldsymbol{Z}|\pi)p(\pi)\\ &\propto \pi^{n_{12}(\boldsymbol{Z})}(1-\pi)^{n_2 - n_{12}(\boldsymbol{Z})}\pi^{\alpha_\pi - 1}(1-\pi)^{\beta_\pi - 1}\\ &\propto \pi^{n_{12}(\boldsymbol{Z})+\alpha_\pi - 1}(1-\pi)^{n_1 - n_{12}(\boldsymbol{Z})+\beta_\pi - 1}.\end{aligned}$$

Thus $\pi^{(s+1)}|\boldsymbol{Z}^{(s+1),\alpha_\pi,\beta_\pi}$ has a Beta$(n_{12}(\boldsymbol{Z}) + \alpha_\pi, n_2 - n_{12}(\boldsymbol{Z}) + \beta_\pi)$ distribution.

Let $\Gamma_{.j}$ denote the set of n_1 comparison vectors with $j \in \boldsymbol{X_2}$. We have

$$\begin{aligned}p(\Gamma_{.j}|Z_j = z_j, \boldsymbol{m}, \boldsymbol{u}) &\propto \prod_{i=1}^{n_1}\left[\prod_{f=1}^{F}\prod_{l=1}^{L_f} m_{fl}^{I(Z_j=i)} u_{fl}^{I(Z_j \neq i)}\right]^{I(\gamma_{ij}^f = l)I_{obs}(\gamma_{ij}^f)}\\ &\propto \prod_{i=1}^{n_1}\frac{\left[\prod_{f=1}^{F}\prod_{l=1}^{L_f} m_{fl}^{I(Z_j=i)} u_{fl}^{I(Z_j \neq i)}\right]^{I(\gamma_{ij}^f = l)I_{obs}(\gamma_{ij}^f)}}{\prod_{f=1}^{F}\prod_{l=1}^{L_f} u_{fl}^{I(\gamma_{ij}^f = l)I_{obs}(\gamma_{ij}^f)}}\\ &\propto \prod_{i=1}^{n_1}\left(\prod_{f=1}^{F}\prod_{l=1}^{L_f}\frac{m_{fl}}{u_{fl}}\right)^{I(z_j=i)I(\gamma_{ij}^f = l)I_{obs}(\gamma_{ij}^f)}\\ &= \begin{cases} w_{z_j,j} & z_j \leq n_1;\\ 1 & z_j = n_1 + j,\end{cases}\end{aligned}$$

where

$$w_{ij} = \prod_{f=1}^{F}\prod_{l=1}^{L_f}\left(\frac{m_{fl}}{u_{fl}}\right)^{I(\gamma_{ij}^f = l)I_{obs}(\gamma_{ij}^f)}.$$

With the likelihood in this form, we derive an expression for the posterior distribution of $\boldsymbol{Z}$. Because we sample each Z_j independently of all other $Z_{j'}$, we use only the full conditional for an individual Z_j:

$$
\begin{aligned}
&p\left(Z_j^{(s+1)} = z_j | \Gamma_{.j}, \boldsymbol{m}, \boldsymbol{u}, \boldsymbol{Z^{(s)}}, \pi\right) \\
&\quad \propto p(\Gamma_{.j} | Z_j^{(s+1)}, \boldsymbol{m}, \boldsymbol{u}) P(Z_j^{(s+1)} | \pi) \\
&\quad \propto \left(\sum_{i=1}^{n_1} w_{z_j,j} I(z_j = i) + I(z_j = n_1 + j)\right) \left(\pi \sum_{i=1}^{n_1} \frac{1}{n_1} I(z_j = i) + (1-\pi) I(z_j = n_1 + j)\right) \\
&\quad = \frac{\pi}{n_1} \sum_{i=1}^{n_1} w_{z_j,j} I(z_j = i) + (1-\pi) I(z_j = n_1 + j). \\
&\quad = \begin{cases} \frac{\pi}{n_1} w_{z_j,j} & z_j \leq n_1; \\ 1 - \pi & z_j = n_1 + j. \end{cases}
\end{aligned}
$$

For more direct comparability with the method from Sadinle (2017), we avoid sampling π directly, and instead integrate over it in the final full conditional for Z_j.

$$
\begin{aligned}
&p\left(Z_j^{(s+1)} = z_j | \Gamma, \boldsymbol{m}, \boldsymbol{u}, \boldsymbol{Z^{(s)}}\right) \\
&\quad = \int_\pi p\left(Z_j^{(s+1)} | \Gamma, \boldsymbol{m}, \boldsymbol{u}, \boldsymbol{Z^{(s)}}, \pi\right) p\left(\pi | \boldsymbol{Z^{(s)}}\right) d\pi \\
&\quad = \int_\pi \left[\frac{\pi}{n_1} w_{z_j,j} I(z_j \leq n_1) + (1-\pi) I(z_j = n_1 + j)\right] p\left(\pi | \boldsymbol{Z^{(s)}}\right) d\pi \\
&\quad = \frac{\int_\pi \pi p\left(\pi | \boldsymbol{Z^{(s)}}\right) d\pi}{n_1} w_{z_j,j} I(z_j \leq n_1) + \left(1 - \int_\pi \pi p\left(\pi | \boldsymbol{Z^{(s)}}\right) d\pi\right) I(z_j = n_1 + j) \\
&\quad = \frac{n_{12}(\boldsymbol{Z}) + \alpha_\pi}{n_1(n_2 + \alpha_\pi + \beta_\pi)} w_{z_j,j} I(z_j \leq n_1) + \frac{n_2 - n_{12}(\boldsymbol{Z}) + \beta_\pi}{n_2 + \alpha_\pi + \beta_\pi} I(z_j = n_1 + j) \\
&\quad \propto w_{z_j,j} I(z_j \leq n_1) + n_1 \frac{n_2 - n_{12}(\boldsymbol{Z}) + \beta_\pi}{n_{12}(\boldsymbol{Z}) + \alpha_\pi} I(z_j = n_1 + j) \\
&\quad = \begin{cases} w_{z_j,j} & z_j \leq n_1; \\ n_1 \frac{n_2 - n_{12}(\boldsymbol{Z}) + \beta_\pi}{n_{12}(\boldsymbol{Z}) + \alpha_\pi} & z_j = n_1 + j. \end{cases}
\end{aligned}
$$

8.3 One Hot Encoding Transformation

As described in Section 4, `fabl` makes use of one-hot encodings to aid in vectorized computations. For γ_{ij}^{f} with L_f levels, define e_{ij}^{f} to be an $L_f \times 1$ vector. When $\gamma_{ij}^{f} = l$, we set the l^{th} element of e_{ij}^{f} to be 1, and set the other $L_f - 1$ elements of e_{ij}^{f} to be 0. We then concatenate the e_{ij}^{f} for all $f \in \{1, \ldots, F\}$, resulting in the one-hot encoded comparison vector e_{ij} of length $\sum_{f=1}^{F} L_f$.

For example, consider comparing the toy records shown in Table 8 with $L = (3, 3, 2, 2)$ levels of agreement for last name, first name, DOB, and city respectively. Since the first name differs by only one letter, a reasonable comparison vector for this pair would be $\gamma_{ij} = (1, 2, 1, 2)$. The one hot encoding representation of this vector is $e_{ij} = (1, 0, 0, 0, 1, 0, 1, 0, 0, 1)$.

Last Name	First Name	DOB	City
Smith	Taylor	01/01/2000	Durham
Smith	Tayler	01/01/2000	Raleigh

Table 8: Example records for one hot encoding.

8.4 Accuracy under Partial Estimates

In this section, we repeat the simulation study in Section 5.2 of the main text, allowing for clerical review rather than forcing all records to have or not have links. Specifically, by leaving $\theta_{10} = \theta_{01} = 1$ and $\theta_{11} = 2$, but setting $\theta_R = 0.1$, we allow the model to decline to decide a match for certain records, with nonassignment being 10% as costly as a false match. In this context, we are no longer focused on finding all true matches, but rather protecting against false matches. Thus, instead of recall, we use the negative predictive value (NPV), defined as the proportion of non-links that are actual non-matches. Mathematically, NPV $= \sum_{j=1}^{n_2} I(\hat{Z}_j = Z_j = n_1 + j) / \sum_{j=1}^{n_2} I(\hat{Z}_j = n_1 + j)$. We continue to use the precision, which is renamed the positive predictive value (PPV) in this context. Lastly, we also examine the rejection rate (RR), or how often the model declines to make a linkage decision, defined as RR $= \sum_{j=1}^{n_2} I(\hat{Z}_j = R)/n_2$. To convey this information alongside NPV and PPV, for which values close to 1 indicate strong performance, we report the decision rate (DR), defined as DR $= 1 - RR$.

In Figure 7, we see that `fabl` maintains equivalently strong PPV as `BRL` across all linkage settings. However, with high amounts of error, and thus fewer accurate and discerning fields of information, the rejection rate under `fabl` rises, leading to a decrease in NPV. Since `fabl` does not remove previously matched records from consideration for a new record, posterior probabilities of matches at times can be split across more records; in contrast, `BRL` is able to maintain higher confidence in matches in this setting. If one wishes to use partial estimates, `fabl` will possibly leave more linkages for the modeller to match by hand than would be left under `BRL`, but the decisions made by each method will have nearly equal accuracy.

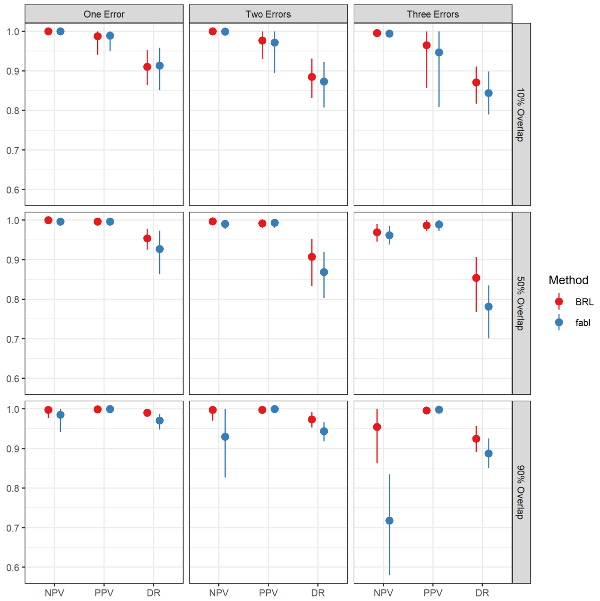

Figure 7: Negative predictive value (NPV), positive predictive value (PPV), and decision rate (DR) on data files in the simulation in Appendix 8.4. We see poorer performance for `fabl` only in situations with high overlap.

8.5 Traceplots for Simulation Study

Below are traceplots for one of the 900 linkage tasks that comprise the simulation in Section 5.2. It is set up with one error across the linkage fields and 50 duplicates across files. Traceplots across other settings exhibit similar behavior. Note that traceplots for $\boldsymbol{u}$ parameters show very little variation because the overwhelming majority of record pairs are nonmatching.

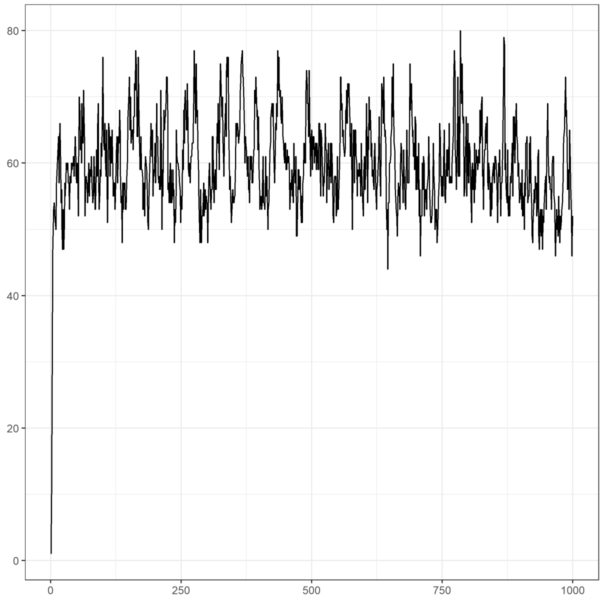

Figure 8: Representative traceplot of overlap between files from simulation study in Section 5.2.

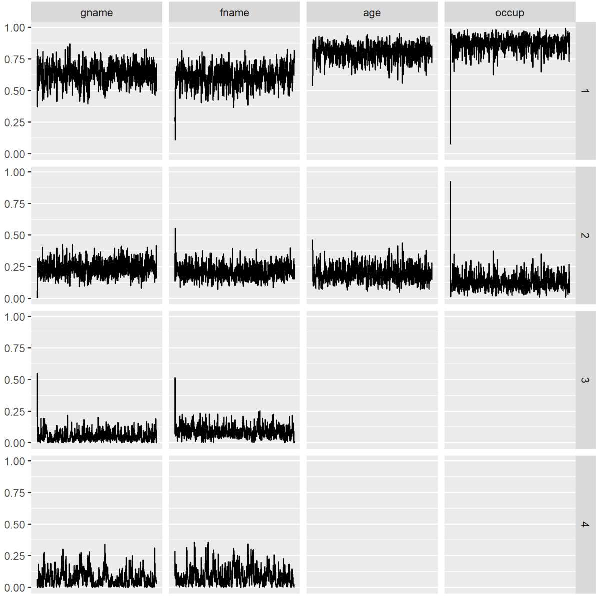

Figure 9: Representative traceplot of $\boldsymbol{m}$ parameter from simulation study in Section 5.2.

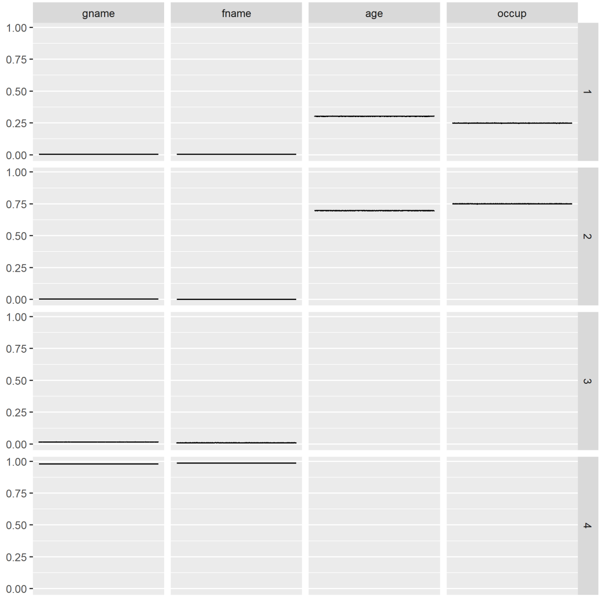

Figure 10: Representative traceplot of $\boldsymbol{u}$ parameters from simulation study in Section 5.2.

	$\boldsymbol{m}$			$\boldsymbol{u}$		
	Agree	Partial	Disagree	Agree	Partial	Disagree
Feature 1	$\frac{9}{10}$	$\frac{9}{100}$	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{3}{100}$	$\frac{96}{100}$
Feature 2	$\frac{9}{10}$	$\frac{9}{100}$	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{3}{100}$	$\frac{96}{100}$
Feature 3	$\frac{9}{10}$	$\frac{9}{100}$	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{3}{100}$	$\frac{96}{100}$
Feature 4	$\frac{9}{10}$	$\frac{9}{100}$	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{3}{100}$	$\frac{96}{100}$

Table 9: Probabilities used for $\boldsymbol{m}$ and $\boldsymbol{u}$ distributions in simulation study in Appendix 8.6.

8.6 Additional Speed Simulation Study

To illustrate that different constructions of the comparison vectors lead to similar speed gains, we replicate the speed study of Section 5.1 under different settings. Here, we use four fields of comparison, each with three possible levels of agreement, resulting in $3^4 = 81$ possible patterns. The $\boldsymbol{m}$ and $\boldsymbol{u}$ parameters for this simulation are shown Table 9.

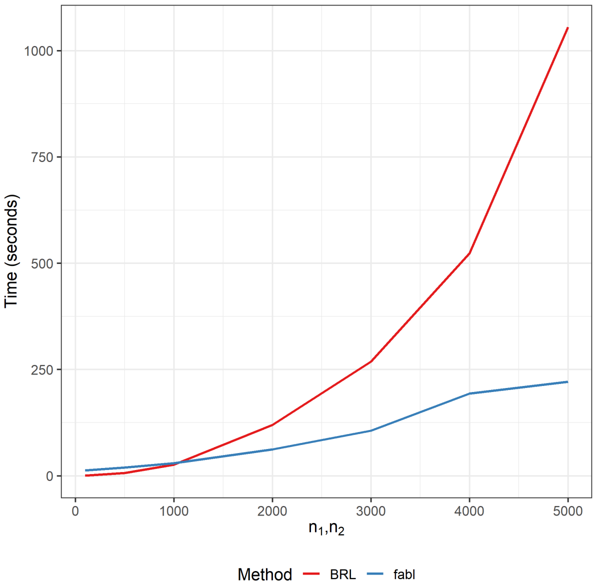

Figure 11: Run-time for `BRL` and `fabl` to run 1000 Gibbs iterations in simulation study in Appendix 8.6, including hashing step for `fabl`, for increasing values of both n_1 and n_2. We see near quadratic growth in run-time for `BRL`, and near linear growth for `fabl`.

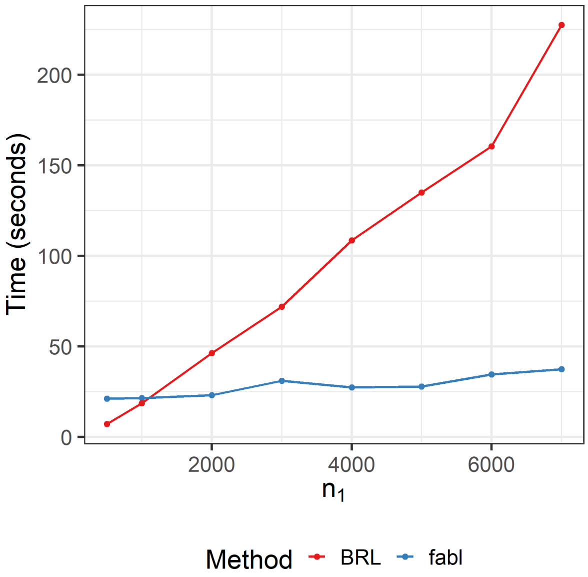

Figure 12: Run-time for `BRL` and `fabl` to run 1000 Gibbs iterations in simulation study in Appendix 8.6, including hashing step for `fabl`, with increasing n_1 and n_2 fixed at 500. We see linear growth in run-time for `BRL`, and near constant run-time for `fabl`.

8.7 Traceplots for El Salvador Case Study

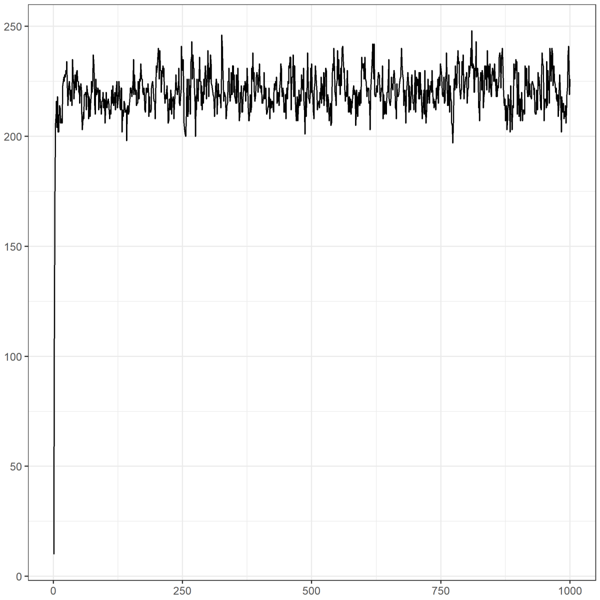

Figure 13: Traceplot for number of matches found across data files in El Salvador case study.

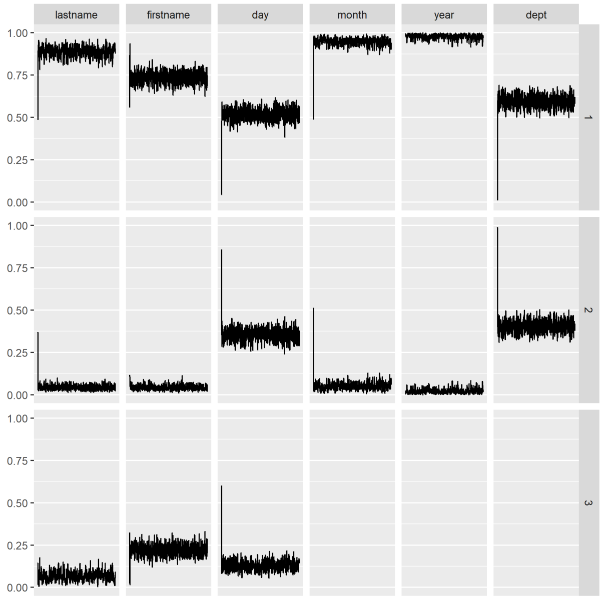

Figure 14: Traceplot for $\boldsymbol{m}$ parameter in El Salvador case study.

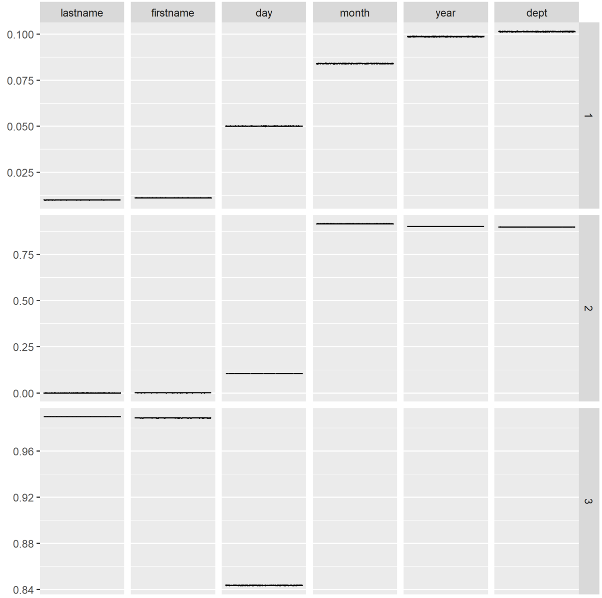

Figure 15: Traceplot for $\boldsymbol{u}$ parameter in El Salvador case study.