Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/120123

Mixed Statistical Matching Approaches Using a Latent Class Model: Simulation Studies

Israa Lewaa[1,*], *Mai Sherif Hafez*[2] *and Mohamed Ali Ismail*[2]

[1] Department of Business Administration, Faculty of Business Administration, Economics and Political Science, The British University in Egypt, Cairo, Egypt
[2] Department of Statistics, Faculty of Economics and Political Science, Cairo University, Cairo, Egypt

Received: 29 Jul. 2022, Revised: 28 Aug. 2022, Accepted: 15 Sep. 2022
Published online: 1 Jan. 2023

Abstract: In the era of data revolution, availability and presence of data is a huge wealth that has to be utilized. Instead of making new surveys, benefit can be made from data that already exists. As, enormous amounts of data become available, it is becoming essential to undertake research that involves integrating data from multiple sources in order to make the best use out of it. Statistical Data Integration (SDI) is the statistical tool for considering this issue. SDI can be used to integrate data files that have common units, and it also allows to merge unrelated files that do not share any common units, depending on the input data. The convenient method of data integration is determined according to the nature of the input data. SDI has two main methods, Record Linkage (RL) and Statistical Matching (SM). SM techniques typically aim to achieve a complete data file from different sources which do not contain the same units. There are a number of traditional matching techniques mentioned in the literature. Among these techniques, there are various approaches for continuous data, but not as many methods for categorical data. This paper proposes a Statistical Matching technique for categorical data based on latent class models within a Bayesian framework. Dirichlet Process Mixture of Product of Multinomial distributions model is used in Statistical Matching throughout this paper which is a fully Bayesian estimation method for latent class models. Performance of the proposed latent class model used for Statistical Matching is evaluated using an empirical comparison with several existing matching procedures based on simulation studies.

Key Words: Bayesian Statistical Matching; Categorical data; Dirichlet process; Latent Class Model; Mixed Methods.

1 Introduction

Attention has been growing towards Data Integration in response to the increasing flood of available data. Data Integration aims at integrating two or more data sources (usually data from sample surveys) sharing the same target population. Statistical Data Integration (SDI) can be used to integrate data files that have common units, and it also allows to merge unrelated files that do not share any common units, depending on the input data. The convenient method of data integration is determined according to the nature of the input data. SDI has two main methods, Record Linkage (RL) and Statistical Matching (SM). [1] provide a review of data integration techniques for combining probability samples, probability and nonprobability samples, and probability and big data samples. [2] presents a review for combining data from different sources focusing on record linkage. RL is the method of gathering information from multiple sources that relates to the same entity. If there is a unique identifier er in the data sources that will be integrated, then there is no difficulty in matching data sources.

*Corresponding author e-mail: Israalewaa@feps.edu.eg

In this case, deterministic linkage will be used. Deterministic linkage is a direct way of linking which usually requires exact agreement on a unique identifier (such as a national identity number). If there is no unique identifier, the Probabilistic Record Linkage is employed (see [3,4,5,6] for details). The advantage of SM is that it can be used as a supplement to RL to reduce the potential bias when obtaining the estimates using RL [7]. On the other hand, SM techniques typically aim to achieve a complete data file from different sources which do not contain the same units. In SM, data sources just share a set of common variables and inference is required on the other variables. Besides the main goal of SM which is data integration, SM also has the objective of jointly analyzing pairs of variables observed in two distinct sample surveys [8,9,10,11]. Inconsistent SM can only produce meaningless data that are impossible to compare and use. Accordingly, suitable approaches should be used for SM to keep pace with rapid development in the data science.

The main goal of Statistical Matching (SM) is data integration. SM also has the objective of jointly analyzing pairs of variables observed in two distinct sample surveys [11,12,7,13,14]. Hence, the general benefit of SM is the creation of a full source of data that contains information about all variables from distinct sources that do not share the same units. This allows better use of data that is already available, making it a cost effective and timely way of gathering more information without the need to collect new data. This can improve decision making and data quality. There are a lot of approaches mentioned in the literature for SM either under Conditional Independence Assumption (CIA) or auxiliary information, along with their drawbacks whenever applicable [15,16,17]. It is noted that available SM procedures focus more on the case of continuous variables, rather than categorical data despite categorical data being the most common type of data in many social surveys. We thus pay more attention to categorical data in this paper by proposing a new Statistical Matching technique particularly designed for categorical data.

Since SM can be viewed as a problem of missing data [18], some methods mentioned in the literature for handling incomplete categorical data are restructured to be used in the context of SM. These methods are Multiple Imputation (MI) using a Log Linear Model (LLM) and Multiple Imputation by Chained Equations (MICE) (see [19, 20, 12, 21] for details). Recently, latent class models (LCM) have been used for multiple imputation. We therefore propose to employ LCMs to be used in the SM framework in the same manner that LLM and MICE are implemented for the same purpose. In the literature, there are four estimation methods used for LCM for MI. These are Maximum Likelihood LCM (MLLC) and Divisive LC model (DLC) which are frequentist methods. Whereas, the standard Bayesian LC model (BLC) and the Dirichlet Process Mixture of Product of Multinomial distributions model (DPMPM) are Bayesian methods.

The main motivation to use LCMs in SM, besides the limited number of existing techniques for SM in case of categorical data, is to overcome the disadvantages and problems arising from the use of LLM and MICE. The validity of the Conditional Independence Assumption, between variables that are not jointly observed, requires the best choice of common variables. [12] Suggests the use of latent classes to create and maintain CIA within classes, when relevant common variables cannot be found. Moreover, the structure of data used in SM can be very complex. In general, this means that the relationship between the variables is more complex than just a linear association, for instance. Another important point to be noted is the high dimensionality in the data due to having a lot of variables. The methods proposed in this paper for SM with categorical variables, use latent class models for MI within a Bayesian framework, taking into consideration the before mentioned problems.

The organization of this paper is as follows: Section 2 gives a review of existing approaches for SM in case of categorical data. Section 3 gives a discussion of the use of LCM for MI. Section 4 presents the proposed model for SM based on Dirichlet Process Mixture of Product of Multinomial distributions model (DPMPM) with a full explanation of the steps of our proposed mixed methods based on CIA and auxiliary information. Furthermore, a comparison of our proposed methods either in case of CIA or auxiliary information with some traditional SM techniques is carried out based on a simulation studies in Section 5. Finally, Section 6 gives conclusion and future points.

2 Reviews of Existing Approaches for Statistical Matching with Categorical Data

There are two main methods in the literature for MI that are used for categorical data and then used for statistical matching. These are Log Linear Model (LLM) and Multiple Imputation by Chained Equations (MICE). MI using LLM, proposed by [22], works on capturing the relevant associations in the joint distribution of a series of categorical variables. SM for categorical data using LLM is presented as a mixed method of two steps. On the other hand, MICE is the second method used for SM of categorical data, introduced by [12]. MICE, a common modification action in MI, are an iterative method that considers the imputation problem as a set of estimations where each variable takes its turn in being regressed on the other variables. MICE is a flexible method that handles various types of variables, since each variable is imputed using its own imputation model [23]. This procedure is called chained equations, variable-by-variable Gibbs sampling, or regression switching. For more details (see [24,12]). MICE can be applied for both continuous and categorical data. [12] found out that MI approaches are superior to traditional SM procedures in case of continuous data, such as hot deck procedures, and

thus suggested to use it as a method for SM.

One major limitation for the use of LLM for MI in general is that it can only be used if the number of variables is relatively limited. Since the processing cells number in the multi-way cross tabulation increases with the variables number exponentially [25], LLM can not be applied in this case of having a large number of variables. This is considered a serious defect in LLM and consequently it is necessary to search for alternative approaches to deal with a large number of variables.

On the other hand, MICE also suffers from drawbacks although it is an intuitive and realistic technique. One of its disadvantages is that there is no theoretical background for the convergence of missing data draws with the resulting distribution of missing data. Secondly, MICE mainly involves the main effects when proceeding a series of regression equations. This may lead to not picking up higher order associations between the variables. In addition, while the approach requires interactions in the higher order to be used, this can be a completely challenging and time consuming process in case of having large number of variables in the imputation process [25].

[26] proposed several mixed procedures in case of categorical data for SM based on multinomial logistic regression models depending on CIA and auxiliary information. Other SM approaches are based on Bayesian networks. [27] described and discussed first attempts for SM of discrete data using Bayesian networks. A further study of how probabilistic graphical models may be used for SM is introduced by [28], who performed log-linear Markov networks under the assumption of conditional independence. For a detailed review of recent approaches and developments of SM techniques, see [15].

An imputation model using LCM is introduced in case of categorical data by [25]. Different problems associated with LLM and MICE appear to be solved by LCM for MI. Even if the variables number is large, LCM can be estimated efficiently [29]. In addition, complex interactions of higher order as well as simple associations between the variables are considered in the imputation process with models that contain a sufficient number of latent classes [30]. Therefore, LCM can be used for datasets drawn from large scale studies, where there is a large number of variables and complex relationship structures [31,32]. It is important to mention that latent variable models have also already been used in the end of record linkage, which is the second type of data integration. For more information about these methods of record linkage, see [33,34,35,36,37]. These methods used a Bayesian approach to graphical record linkage and duplication. [38] presented an approach using Dirichlet Process Mixture of Product of Multinomial distributions model (DPMPM) for imputing the missing values in case of categorical variables through NPBayesImputeCat package in R.

In our study, we utilize this same model to be used in the context of SM. We present a new mixed approach, in which, we have two steps. The first step is to perform DPMPM and the second step is the hotdeck method, which is the same manner of previous approaches provided for SM such as loglinear method. Moreover, we will compare our new SM mixed method with existing SM methods. To extend this comparison, the method proposed by [38], which is demonstrated for imputing the missing data, will be exploited in SM. This comparison will be conducted via simulation studies in different scenarios in both cases, under conditional independence assumption and with auxiliary information.

3 Latent Class Model for Multiple Imputation

This section aims to present an overview of MI by employing a LCM to overcome difficulties of the existing approaches in SM in case of categorical data. The four various estimation methods of LCM for MI will be considered. LCM is a type of mixture model which describes the distribution of categorical data [39, 40 y].

Besides using them for data reduction, latent class models have been recently used for MI where they are employed as a tool for estimating the density of categorical variables [25,41]. An LCM can be estimated efficiently even if the number of variables is large [29,32,42], making it suitable for clustering or density estimation for datasets drawn from large-scale studies, where there is a large number of variables and complex relationship structures.

By assuming that we have a fixed number of latent classes K, we can de ne some notations as follows;

$z_i \in \{1,\dots,K\}$ is the class membership of unit i for $i = 1,\dots,n$,

$\Psi z = (z_i = k)$ is the probability of belonging to class k, where; $\Psi = (\Psi_1,\dots,\Psi_\infty)$ is the same for all units,

$\Phi_{y|z} = r(y_{ij} = y \mid z_i = k)$ is the conditional probability of $y_{ij} = y$, given that unit i is in class k for any value, where;

$\Phi = \{\Phi_{y|z} : y = 1,\dots,t_j, j = 1,\dots,p, k = 1,\dots,K\}$ is the collection of all $\Phi_{y|z}$.

The posterior class membership probabilities of the units are a quantity of concern while using LCM. This quantity represents the probability that a unit belongs to the k^{th} class given the observed data pattern y_i. Bayes' theorem can describe this quantity in the following manner:

$$P(z_i = k|\mathrm{y_i}) = \frac{P(z_i=k)P(\mathrm{y_i}|z_i = k)}{P(\mathrm{y_i})} \tag{3.1}$$

When a LCM is employed for MI in case of data with missingness, there is no need to de ne meaningful clusters, but the joint density of variables in the imputation model should be well defined. Using LCM for imputation purposes is being a device for the estimation of (y_i). Once an incomplete dataset is used to estimate a LCM, MI can be done by M draws of imputations in a random way from the posterior distribution of the missing values for each non-response, given the observed data and model parameters. Through MI model, the missing data are replaced (or predicted, imputed) $M > 1$ times by different values, the distribution of which is estimated with the imputation model. Hence, the uncertainty of the imputation is considered by repeating this process $m > 1$ times for each unit with at least one missing value. In this case, data is divided into two parts $y_{i,bs}$ and $y_{i,mis}$. $y_{i,bs}$ represents the observed part for unit i, whereas $y_{i,mis}$ is the missing part for unit i. The role of the imputation model, specifically, is to provide values sampled from $P(y_{i,mis}|y_{i,obs})$, that is, the distribution of the missing data given the observed data. In LCM for incomplete data, the conditional distribution $P(y_{i,mis}|y_{i,obs})$ can be written as;

$$P(y_{i,mis}|y_{i,obs}) = \sum_{k=1}^{K} P(z_i = k, y_{i,mis}|y_{i,obs}) \quad (3.2)$$

$$= \sum_{k=1}^{K} P(z_i = k|y_{i,obs})\, P(y_{i,mis}|z_i = k) \quad (3.3)$$

$$= \sum_{k=1}^{K} P(z_i = k|y_{i,obs}) \prod_{j=1}^{p} [P(y_{ij}|z_i = k)]^{1-r_{ij}}. \quad (3.4)$$

Where;
r_{ij} =0 if y_{ij} is missing and 1 if observed.

$$P(z_i = k|y_{i,obs}) = \frac{P(z_i = k)P(y_{i,obs}|z_i = k)}{P(y_{i,obs})}$$

$$P(\mathbf{y_{obs}}) = \sum_{k=1}^{K} P(z_i = k) \prod_{j=1}^{p} [P(y_{ij}|z_i = k)]^{r_{ij}}$$

Equation (3.2) is obtained using the local independence assumption between $y_{i,bs}$ and $y_{i,mis}$ given z_i. Whereas equation (3.3) uses again the local independence assumption, but among the variables containing missing values.
There are four various estimation methods of LCM for MI. These are Maximum Likelihood LCM (MLLC) and Divisive LC model (DLC) which are frequentist methods. Whereas, the standard Bayesian LC model (BLC) and the Dirichlet Process Mixture of Product of Multinomial distributions model (DPMPM) are Bayesian methods. These four models can be used as a perfect MI tool with large data when variables are categorical [29,43,44]. Since SM can be viewed as a special case of imputation, we will use one of these four estimation methods in the context of SM. Namely, we choose to implement DPMPM for the following reasons:

- [45] Stated that the validity of the imputation inferences needs analysts for incorporating all uncertainty causes, including the estimation of the parameter. Since this uncertainty can be incorporated immediately within Bayesian estimation, we choose DPMPM, a fully Bayesian estimation method, where the number of classes is determined within the estimation process.

- When the relationship between the categorical variables is more complex than just a linear association with missing data, there is a need for a MI approach that (i) prevents the difficulties inherent in LLM concerning model selection and estimation, (ii) has theoretical foundations as in a consistent Bayesian joint model, and (iii) offers efficient computation. All of these advantages are available in DPMPM.

- It takes structural zero into consideration. Structural zero is an important feature of survey data, defined by the existence of impossible combinations of variables. For example, in the combinations of variables of pregnancy status and gender, there should not exist a pregnant male. For household survey, in the combinations of variables of relationship and age, there should not exist a household where a son is older than his biological father, and so on. DPMPM takes structural zeros into consideration by assigning zero probability for impossible combinations, which is a challenging task.

4 Proposed Mixed Approaches for Statistical Matching Using DPMPM

In this section, we will present a number of mixed methods for SM based on DPMPM for both CIA and auxiliary information. We first give a detailed description for DPMPM model for MI and show the mechanism of this model and how it works before presenting our mixed methods.

In general, the Dirichlet process is a stochastic process used in Bayesian nonparametric models of data, particularly in Dirichlet process mixture models (also known as infinite mixture models). Hence, DPMPM is considered to be a generalization of the finite mixture model for multinomial data, presented in Section 3, in which DPMPM assumes infinite classes number instead of a fixed number of classes. Moreover, DPMPM provides a complete Bayesian modelling method in case of categorical data with high dimensionality of the variables, estimated using Gibbs sampler.

The finite mixture model assumes that a fixed number of latent classes K, but it is not known in practical analysis and should be determined by the researcher. If the researcher puts a too small value for K, the mixture model may not be versatile enough to predict the complicated dependencies. [25] propose that researchers pick K based on penalized likelihood statistics, such as the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC). The efficiency of penalized likelihood model selection processes is not properly studied in this context. Therefore, it is not certain that the AIC (or BIC) chooses a sufficiently large K [29]. And if we suppose that AIC gives a relatively large K, the uncertainty about K when producing the imputations is still ignored until determining the K value. This contradicts the recommendation by [45] to take into consideration all potential uncertainty regarding the imputation model parameters to prevent underestimating the variances of the model parameters [29]. DPMPM deals with the issue of parameter uncertainty, and tackles the problem of choosing a specific measure to determine a fixed K. This problem can be solved by assuming that theoretically there is an in finite number of classes ($K = +\infty$) and this matches with the theoretical background of DPMPM as it is basically an infinite mixture of products of multinomial distributions.

For DPMPM, the likelihood of the data can be given by

$$y_{ij}|z_i, \Phi \sim Multimnomial\,(\Phi_{z_i j1}, \dots, \Phi_{z_i jt})\ \forall\, i, j \tag{4.1}$$

$$z_i|\Psi \sim Multimnomial\,(\Psi_1, \dots, \Psi_\infty)\ \forall\, i \tag{4.2}$$

Where parameters prior distributions are assumed as follows

$$\Phi_{y|z} = (\Phi_{kj}, \dots, \Phi_{kjlj}) \sim Dirichlet\,(a_{j1}, \dots, \Phi_{jtj}) \tag{4.3}$$

$$\Psi_z = V \prod_{g<k}(1 - V_g);\ k = 1, \dots, \infty \tag{4.4}$$

$$V_k \sim Beta\,(1, a) \tag{4.5}$$

$$a \sim \text{Gamma}\,(\alpha, \beta) \tag{4.6}$$

A stick-breaking procedure is one of the potential principles of the Dirichlet process. It acts as a prior for the mixture probabilities Ψ_z [46,47]. Therefore, the prior distribution is modelled for the mixture probabilities, $\Psi z = \Pr(\Psi_1, \dots, \Psi_\infty)$, by using the stick-breaking representation of the Dirichlet process as in equations 4.4 and 4.5.

The joint posterior distribution of the parameters can be approximated via MCMC. According to the system of equations mentioned above, V_k is drawn from a Beta distribution with parameters $(1, a)$ for $k = 1, \dots, \infty$; where a follows a $Gamma\,(\alpha, \beta)$. In this context, [48] suggest the parameters of Gamma distribution to be α = 0.25, β = 0.25. This makes each V_k to be uniformly distributed in the (0,1) range, whereas each element of $(a_{j1}, \dots, a_{jtj})$ equalized to one to correspond to a uniform distribution. For adapting with missing data, truncation process is considered. Truncation of the stick breaking probabilities Ψz at an arbitrary large K^* is suggested by [29] due to the difficulty to deal with infinite number of classes K practically. They just put a condition that K^* not to be too large to guarantee the fast computing. After that, Gibbs sampler is conducted. This requires running an algorithm containing six main steps: the first is to assign each unit to a latent category by sampling draws from the posterior membership probabilities ($z_i = k|y_i$) in equation 3.1. Secondly, because of the truncation, some updates occur to the system of equations mentioned above and the sample draws for them are as follows;

$$V_k \sim Beta\,(1 + v, a + \theta);\ \mathrm{k} = 1, \dots, K^* - 1 \tag{4.7}$$

where;
ν is the number of units allocated in the kth latent class,
θ is the number of units assigned to the latent classes which go from k + 1 to K^*

For the third step in the algorithm, we set $V_{K*} = 1$ and each Ψz is calculated through the formula $\Psi z = V_k \prod_{g<k} (1 - Vg)$. In the fourth step, draw $\Phi_{y|z}$ from

$$\Phi_{y|z} = (\Phi_{kj}, \ldots, \Phi_{kjlj}) \sim Dirichlet(a_{j1} + \omega, \ldots, a_{jtj} + \omega) \quad (4.8)$$

where;
ω is the number of units, which take one of the possible observed values of the j^{th} variable and are dropped into the k^{th} latent class.
In the fifth step, a is drawn from Gamma distribution as follows;

$$a \sim Gamma(\alpha + K^* - 1, \beta - log(\Psi_z^{K^*}) \quad (4.9)$$

The last step is the imputation process, given that the value $z_i = k$ of each unit, y_{ij} is sampled by;

$$y_{ij} \mid z_i, \Phi \sim Multinomi(\Phi_{zij1}, \ldots, \Phi_{zijtj}) \quad \forall i, j \quad (4.10)$$

Steps 1-6 are repeated until convergence is attained.
Next, we introduce a number of mixed methods for SM using DPMPM in case of categorical data for both under CIA and with auxiliary information.

4.1 Statistical Matching Methods under CIA

Mixed DPMPM Approach in case of CIA (MDC1)

In the basic SM structure, there exist two files say A and B as sources of data. Now, suppose file A contains variables *X* and *Y*, while file B contains variables *X* and *Z*. In this case, variables *Y* and *Z* are not found jointly in one dataset. In this method, file A and file B will be combined together and the resulting missing pattern is known as Missing Completely At Random (MCAR) by design (see Table 1 in Appendix A). SM aims to provide a complete file containing variables *X, Y* and *Z* by imputing the missing *Z* in file A to get an integrated file A. So, our first proposed method, which we call MDC1, takes the following steps:

1. MDC1 procedure begins by DPMPM to fill the missing data. After running DPMPM step, file A consists of $X_{Observed}^{(A)}$, $Y_{Observed}^{(A)}$ and $\hat{Z}_{Primary}^{(A)}$. File B consists of $X_{Observed}^{(B)}$, $Z_{Observed}^{(B)}$ and $\hat{Y}_{Primary}^{(B)}$.
2. Matching step for MDC1 is done by hot deck method. A live z value $Z_{live}^{(B)}$ that is observed in file B is assigned to the a^{th} record in file A according to the nearest distance $d_{ab}((Y_{Observed}^{(A)}, \hat{Z}_{Primary}^{(A)}), (\hat{Y}_{Primary}^{(B)}, Z_{Observed}^{(B)})$.

Mixed DPMPM Approach in case of CIA (MDC2)

1. The first step is same as MDC1.

2. For each unit in file A, $\widehat{\Phi}_j^{(A)}$ is estimated by fitting a multinomial logistic regression model in which the common variables $X_{Observed}^{(A)}$ and $Y_{Observed}^{(A)}$ are the independent variables and the primary variable $\hat{Z}_{Primary}^{(A)}$ in file A is the dependent variable.

3. For each unit in file B, $\hat{\phi}_j^{(B)}$ is estimated by fitting a multinomial logistic regression model in which the common variables $X_{Observed}^{(B)}$ and $\hat{Y}_{Primary}^{(B)}$ are the independent variables and the observed variable $Z_{Observed}^{(B)}$ in file B is the dependent variable.
4. Distance hot deck approach is applied in which a value of *Z* in file B is imputed for file A based on the nearest distance $d_{ab}\left(\hat{\phi}_j^{(A)}, \hat{\phi}_j^{(B)}\right)$.

Mixed DPMPM Approach in case of CIA (MDC3)

1-2. The first two steps are the same as MDC2.

3. Obtain categorical variable *Z* for the recipient file A by generating multinomial random variable with the predicted probability $\hat{\phi}_j^{(A)}$ in step (2).

4.2 Statistical Matching with Auxiliary Information

Mixed DPMPM Approach in case of Auxiliary Information (MDA1)

1. MDA1 procedure begins by running DPMPM using file A, file B and file C.

2. Matching step for MDA1 is done by hot deck method. A live z value $Z_{live}^{(C)}$ that is observed in file C is then assigned to the a^{th} record in file A according to the nearest distance $d_{ab}((Y_{Observed}^{(A)}, \hat{Z}_{Primary}^{(A)}), (Y_{Observed}^{(C)}, Z_{Observed}^{(C)})$.

Mixed DPMPM Approach in case of Auxiliary Information (MDA2)

1. The first step is same as MDA1.

2. For file A, t a multinomial logistic regression model in which $\hat{Z}_{Primary}^{(A)}$ is the dependent variable and $X_{Observed}^{(A)}$ *and* $Y_{Observed}^{(A)}$ is the independent variables to get probability $\hat{\phi}_j^{(A)}$.
3. For file C, fit a multinomial logistic regression model in which $Z_{Observed}^{(C)}$ is the dependent variable and $X_{Observed}^{(C)}$ *and* $Y_{Observed}^{(C)}$ is the independent variables to get probability $\hat{\phi}_j^{(C)}$.
4. Distance hot deck approach is applied in which a value of *Z* in file C is imputed for file A based on the nearest distance $d_{ab}\left(\hat{\phi}_j^{(A)}, \hat{\phi}_j^{(C)}\right)$.

Mixed DPMPM Approach in case of Auxiliary Information (MDA3)

1. The first step is same as MDA1.

2. Using file B, fit a multinomial logistic regression model in which $Z_{Observed}^{(B)}$ is the dependent variable and $X_{Observed}^{(B)}$ and $\hat{Y}_{Primary}^{(B)}$ are the independent variables.

3. Using file A, the predicted probability $\hat{\phi}_j^{(A)}$ is obtained by using the following equation;

$\hat{\phi}_j^{(A)} = \frac{\exp(\beta_j' X)}{1+\sum_{i=1}^{J-1}\exp(\beta_j' X)}$, j=1,…,J-1. Where β_j' is estimated from step (2).

4. Obtain categorical variable *Z* for the recipient file A by generating multinomial random variable with the predicted probability $\hat{\phi}_j^{(A)}$ in step (3).

5 Simulation Studies

To compare the performance of the proposed SM methods in Section 4 with loglinear, MICE and hot deck, simulation studies are performed. Separate simulation studies are carried out under CIA and with auxiliary information. The effect of

different sample sizes of auxiliary file C is also studied, via another simulation study.

In this section, we present the three steps of our simulation studies. These are data generation, simulation design, and simulation results.

5.1 Data Generation

For our simulation studies, we generate a population of size 50,000 with a number of categorical variables (*X,Y,Z*). Three common variables; X_1, X_2, and X_3, are generated with 4, 3, and 2 categories, respectively. The *Y* and *Z* variables, which are not found jointly found in the same file, we are generated with 3 and 4 categories, respectively where *Y* is found in file B, and *Z* is found in file A. The choice of number of common variables and number of categories here is arbitrary. Through the simulation process, file A, the recipient file, and file B, the donor file, are created using random sampling. In each replication, recipient file A is created by randomly selecting 1500 units of the generated data, while variable *Z* is being removed. Similarly, file B is generated of size 5000 with *Y* being removed. Accordingly, *X* and *Y* are observed in the recipient file A. While *X* and *Z* are observed in the donor file B. This process is repeated 1000 times in order to obtain 1000 different sets of files A and B. The integrated file A will be used to evaluate the performance of our proposed SM methods, by comparing the resulting integrated data to the true generated complete dataset (*X*, *Y*, *Z*), using a relevant measure.

5.2 Simulation Design

The generation of *X*, *Y* and *Z* is performed basically using multinomial logistic regressions in order to consider different associations between variables. *X* variables are created from multinomial distribution with probabilities shown in table 2 in Appendix A. The vector of common variables, $X = (X_1, X_2, X_3)'$, includes three categorical variables with 4,3, and 2 categories, respectively. This vector is converted into dummies as,

$$X_d = (x_{1d,}\, x_{2d,3d})' = (x_{11},\, x_{12},\, x_{13,}x_{14,}x_{21},\, x_{22},\, x_{23,}x_{31,}x_{32})',$$

where $x_{ij} = 1$ if the observation belongs to j^{th} category of the i^{th} categorical variable, and 0 otherwise.

Now, we generate Y based on a multinomial logistic regression model to obtain a certain association between *X* and .

Assume that variable *Y* has *J* categories and that the probability of belonging to each category is ($\Phi_1,\Phi_2,\ldots,\Phi_J$), where $\sum_{j=1}^{J} \phi_j = 1$. Under the multinomial logistic regression model, the log of probability of the observation belonging to each category *j* relative to the last category *J* is

$$ln\frac{\phi_j}{\phi_J} = \beta_j' X, j = 1, \ldots, J-1 \qquad (5.1)$$

where $X = (x_1, x_2, \ldots, x_p)'$ is a vector of explanatory variables and $\beta_j' = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jp})$ is a vector of regression coefficients corresponding to outcome *j*. From Equation 5.1, we get the probability Φ_j as

$$\phi_j = \phi_J \exp(\beta_j' X) = \frac{\exp(\beta_j' X)}{1 + \sum_{i=1}^{J-1} \exp(\beta_j' X)}, \quad j = 1, \ldots., J-1 \qquad (5.2)$$

Y is generated according to the above probabilities, which are functions of the common variables *X*. The variable *Y* is generated having three categories such that $Y_d = (y_1, y_2, y_3) \sim multinomial(\phi_1^y, \phi_2^y, \phi_3^y)$ multinomial where y_1,y_2 and y_3 are dummy variables and ϕ_1^y, ϕ_2^y and ϕ_3^y are the corresponding probabilities of belonging to each category (probability that $y_j = 1$), such that

$$ln\frac{\phi_j^y}{\phi_3^y} = \beta_j'^y X_d, \quad j = 1,2$$

The following values are assumed for parameters in the above model for the generation of the Y variable:

$$\beta_1' = (-0.1, 0.1, -0.1, -0.1, 0.1, -0.1, 0.1, 0.1, -0.1),$$

$$\beta_2' = (-0.1, 0.1, -0.1, -0.1, 0.1, -0.1, 0.1, 0.1, -0.1).$$

Then, Z is generated given the generated X and Y . Ten Z variables are created. Each generated Z having different levels of association with X and Y is included in the dataset (one at a time). Various association strengths of (X,Y,Z) are considered by obtaining 10 different Z variables to evaluate the performance of the different matching methods, under different scenarios of association among variables. Z_1 to Z_5 are generated so that no association exists with X but different level of association exists with Y . Z_6 to Z_{10} are generated so that association exists with X and different level of association exists with Y. Z is generated with four categories from the model

$$ln\frac{\Phi_h^{z_k}}{\Phi_4^{z_k}} = \beta_h'^{z_k}X_d + \lambda_h'^{z_k}Y_d \qquad (5.3)$$

Such that $Z_d = (z_1, z_2, z_3, z_4) \sim multinomial(\phi_1^y, \phi_2^y, \phi_3^y) \sim multinomial(\phi_1^z, \phi_2^z, \phi_3^z, \phi_4^z)$ where z_1, z_2, z_3, z_4 are dummy variables and $\phi_1^z, \phi_2^z, \phi_3^z$ and ϕ_4^z are the corresponding probabilities of belonging to each category (probability that $z_j = 1$), such that $ln\frac{\Phi_h^{z_k}}{\Phi_4^{z_k}} = \beta_h'^{z_k}X_d + \lambda_h'^{z_k}Y_d, h = 1,2,3.$ Z is generated with H categories, $(h = 1,2, H-1)$ and different scenarios (k = 1,2,...,10), and Y_d is a vector of 3 dummy variables.

For generating different scenarios of Z, coefficients are provided in Table 4 in Appendix A. Table 3 in Appendix A shows different levels of association between Y and Z according to Cramer's V measure of association for nominal variables. Values were chosen arbitrarily, however, a variety of values was chosen to capture different levels of association.

Finally and with this simulation design, 1000 completed synthetic datasets are imputed for each one of the 10 scenarios of Z using different SM methods. In other words, for each one of the 10 different scenarios of Z, 1000 datasets are imputed for each method.

5.3 Simulation Results

5.3.1 Simulation results for SM under CIA

To evaluate the performance of the proposed statistical matching methods versus existing methods, we consider the estimation of contingency table of Y and Z using the matched data, since inference on categorical variables is usually based on contingency tables. A set of similarity/dissimilarity measures are computed for marginal or joint distribution of categorical variables. Examples of similarity measures include overlap between two variables and bhattacharyya coefficient, a measure that is based on the difference between frequencies of a contingency table that is constructed from the matched file, and those of the contingency table of the population. Both these measures range within the interval [0,1]. The closer to 1, the more the distributions of variables are similar. Measures of dissimilarity such as total variation distance and Hellinger's distance, also range within the interval [0,1]. The closer to 0, the more the distributions are similar. For details about various distance measures, see [49]. Among these similarity /dissimilarity measures, we use the Hellinger's distance (Hell) for the comparison of the two distributions, which takes the following formula:

$$Hell = \sqrt{1 - Bhatt}$$

Where

$Bhatt = \sum_{m=1}^{M}\sqrt{B_{1,m} - B_{2,m}}$ is the Bhattacharyya coefficient such that $B_{1,m}$ is a relative frequency of $Y \times \hat{Z}$ contingency table in the matched file, and $B_{2,m}$ is a relative frequency of $Y \times Z$ contingency table in the population, and M is number of cells in the $Y \times Z$ contingency table.

We evaluate the performance of our proposed methods by comparing them with LLM, MICE, Random hot deck and other two proposed methods by [26]. We denoted two methods by Kim and Park methods as KP1 and KP2 respectively. Also, a DPMPM method by [38] will be exploited in SM context and denoted by MVR.

The comparison between the different approaches of SM are performed in terms of Hellinger distance for all methods under consideration (Hotdeck, Loglinear, MICE, KP1, KP2, MDC1, MDC2, MDC3) under the ten different simulated scenarios of associations outlined above. Table 5 in Appendix A shows the mean, maximum and minimum of 1000 values of

Hellinger distance. The results indicate that MICE consistently has the worst performance among the eight methods. This may be due to the fact that MICE approach is basically designed for missing data not as a SM technique. Accordingly, MICE is removed from Figure 1in Appendix B that compares different methods of SM, for the sake of a clearer graph. Due to its bad performance, whenever it is included in the graph, the differences among other methods is not visible. On the other hand, all existing methods perform poorly under scenarios where CIA does not hold in the population ($Z4, Z5, Z9, Z10$). On the contrary, our proposed methods (MDC1, MDC2, MDC3) outperform other methods by having the least values for Hellinger distance in case of Z_4, Z_5, Z_9 and Z_{10}. These are scenarios where CIA does not hold in the population. This is because latent class model has the advantage of being based on conditional independence assumption, that is, the scores of different items are independent of each other given latent classes. The results indicate that whenever CIA holds, namely in cases $Z1, Z2, Z3, Z6, Z7, and\ Z8, KP1\ and\ KP2$ exhibit the best performance among all methods.

5.3.2 Simulation Results for SM with Auxiliary Information

In the second simulation, we compare the performance of the statistical matching methods using the auxiliary information suggested in Section 4.2 to the random hot deck method, LLM, MICE and Kim and Park method. To obtain auxiliary information, file C of size 200 with (*X,Y,Z*) variables, about 10% of the recipient file A, was sampled from the population. As in the previous simulation, the mean of 1000 values of Hellinger are shown in Table 6 in Appendix A.

Table 6 in Appendix A shows that in all cases from Z1 to Z10, our proposed SM techniques MDA1 and MDA3 outperform all other methods in which auxiliary information is incorporated, when estimating the joint distribution of (Y,Z), based on the average of Hellinger's distances calculated from 1000 simulations. The performance of MICE has a great improvement when auxiliary information is applied compared with its performance under CIA but it still has a bad performance among all approaches. As expected, the statistical matching method which is valid under CIA is sensitive to the level of association between *Y* and *Z*, while the methods using auxiliary information are relatively insensitive compared to the CIA. Figure 2 in Appendix B summarizes the performance for the other approaches (excluding MICE) existing in Table 6 in Appendix *A*.

5.3.3 Simulation Results for Different Sizes of Auxiliary File

Since MDA1 was found to have the best performance consistently among all methods, in the third simulation we compare the performance of MDA1 with various sample sizes of file C. We choose to compare the performance of the MDA1 method when the size of file C is 50, 100, 150, 200, 250, and 300.

Figure 3 in Appendix B shows that, the performance of MDA1 in estimating the joint distribution of (*Y,Z*) improves as the size of file C increases. See also Table 7 in Appendix A. However, the amount of gain obtained by increasing the size of file C decrease as the size of file C increase which is clearly obvious from figure 3 in Appendix B. Accordingly, almost no significant gain is obtained by increasing n = 150 to n = 300.

6 Conclusions and Future Work

The main goal of a statistical micro matching is to generate synthetic data which combines variables (*Y,Z*) that are not jointly found in the same data file and estimate the joint distribution of (*Y,Z*) and (or) (*X,Y,Z*). Throughout this paper, we introduced a number of methods based on Latent Class Models (LCM) to be used in the context of SM. we proposed several mixed matching methods using DPMPM for categorical variables. First, we proposed a statistical matching method without auxiliary information under CIA. Simulation studies showed that the performance of MICE is the worst among the eight methods under consideration. Unlike our proposed methods, which depend on latent class models, all existing methods performed poorly when CIA did not hold in the population. This is because latent class model has an advantage of being based on an assumption of conditional independence, that is, the scores of different items are independent of each other given latent classes. Whereas in cases where CIA holds, Random method had the best performance. Our proposed mixed method outperforms other methods in cases where CIA does not hold, and is thus more reliable. Second, we proposed a statistical matching method with auxiliary information when CIA does not hold. The performance of MICE had a great improvement when auxiliary information is applied compared with its performance under CIA but it still had a poor performance compared to other approaches. As expected, statistical matching methods that are valid under CIA are sensitive to the level of association between *Y* and *Z*, while the methods using auxiliary information are relatively insensitive compared to the CIA. Finally, the simulation result shows that the size of file C does not need to be large which means the cost to overcome the CIA would not be a serious concern in practice.

As an application to the proposed mixed DPMPM approach, presented in this paper, the authors attempted to integrate data

from two real national surveys, namely the Egyptian Household Income Consumption Expenditure Survey (HICES), carried out by the Central Agency for Public Mobilization and Statistics, and the Egyptian Demographic Health Survey (EDHS), carried out by USAID. The proposed statistical matching technique was used to integrate the two national surveys, thus resulting in a combined dataset that includes data about income, together with data about domestic violence, a combination that has never been available beforehand which allows for the study of various inter-relationships. The quality of the resulting integrated data was validated and results were proved to show good performance. Details about this application are available in [50], currently under review.

For future work, one may consider other types of LCM besides DPMPM and compare their performance within statistical matching. Many other applications can be considered to integrate surveys, applying different SM techniques.

References

[1] S. Yang, J. K. Kim. Statistical Data Integration in Survey Sampling: A Review. arXiv preprint arXiv:200103259. 2020.

[2] M. E. Thompson. Combining data from new and traditional sources in population surveys. International Statistical Review. 2019;87:S79 S89.

[3] I. P. Fellegi and A. B. Sunter. A Theory for Record Linkage. Journal of the American Statistical Association. 1969;64(328):1183 1210.

[4] A. Sayers, Y. Ben-Shlomo, A. Blom and F. Steele. Probabilistic record linkage. International Journal of Epidemiology. 2016 6;45(3):954 64. The Author 2015; Published by Oxford University Press on behalf of the International Epidemiological Association.

[5] J. Murray. Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering. Journal of Privacy and Con dentiality. 2015;7.

[6] M. H. Hof, A. C. Ravelli and A. H. Zwinderman. A Probabilistic Record Linkage Model for Survival Data. Journal of the American Statistical Association. 2017;112(520):1504 1515. Available from: https://doi.org/10.1080/01621459.2017.1311262.

[7] J. Gessendorfer, J. Beste, J. Drechsler and J. W. Sakshaug. Statistical matching as a supplement to record linkage: a valuable method to tackle non-consent bias. Journal of official statistics. 2018;34(4):909 933.

[8] S. Rassler and K. Fleischer. Aspects concerning data fusion techniques. In: International Workshop on Household Survey Nonresponse. vol. 4. DEU; 1998. p. 317 333.

[9] S. Rassler, Fleischer K. An evaluation of data fusion techniques. In: Proceedings of Statistics Canada Symposium 99 on Combining Data from Different Sources; 1999. p. 129 136.

[10] M. D'Orazio, M. Di Zio and M. Scanu. Statistical matching and official statistics. Rivista di Statistica Ufficiale. 2002.

[11] M. D'Orazio, M. Di Zio and M. Scanu. Statistical matching: Theory and practice. John Wiley & Sons; 2006.

[12] S. Rassler. Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches. vol. 168. New York: Springer-Verlag; 2002.

[13] C. Moriarity and F. Scheuren. Statistical matching: a paradigm for assessing the uncertainty in the procedure. Journal of Official Statistics. 2001;17(3):407.

[14] L. C. Zhang . On proxy variables and categorical data fusion. Journal of Official Statistics. 2015;31(4):783 807.

[15] I. Lewaa, M. S. Hafez and M. A. Ismail. Data Integration Using Statistical Matching Techniques: A Review. Statistical Journal of the IAOS. 2021.

[16] P. L. Conti, D. Marella and M. Scanu. Uncertainty analysis in statistical matching. Journal of Official Statistics. 2012;28(1):69

88.

[17] M. D'Orazio, M. Di Zio and M. Scanu. Statistical matching for categorical data: Displaying uncertainty and using logical constraints. Journal of Official Statistics. 2006;22(1):137.

[18] J. k. Kim, E. Berg and T. Park. Statistical matching using fractional imputation. arXiv preprint arXiv:151003782. 2016.

[19] A. C. Singh, G. E. Lemaitre and J. Armstrong. Statistical matching using log linear imputation. Social Survey Methods Division, Statistics Canada; 1988.

[20] A. Singh, H. Mantel, M. Kinack and G. Rowe. Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. Survey Methodology. 1993;19(1):59 79.

[21] S. Rassler. Data Fusion: Identification Problems, Validity, and Multiple Imputation. Austrian Journal of Statistics. 2004 Apr;33(1 2):153 171.

[22] J. L. Schafer. Analysis of Incomplete Multivariate Data. London: Chapman and Hall; 1997.

[23] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas and H. Hemingway. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. American journal of epidemiology. 2014;179(6):764 774.

[24] S. Van Buuren and k. Oudshoorn. Flexible multivariate imputation by MICE. Leiden: TNO; 1999.

[25] J. K. Vermunt, J. R. Van Ginkel, L.A. Van der Ark and K. Sijtsma. Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis. Sociological Methodology. 2008;38(1):369 397.

[26] K. Kim and M. Park. Statistical micro matching using a multinomial logistic regression model for categorical data. Communications for Statistical Applications and Methods. 2019;26(5):507 517.

[27] E. Endres and T. Augustin. Statistical matching of discrete data by Bayesian networks. In: Conference on Probabilistic Graphical Models. PMLR; 2016. p. 159 170.

[28] E. Endres and T. Augustin. Utilizing log-linear Markov networks to integrate categorical data les. 2019.

[29] Y. Si and J.P. Reiter. Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. Journal of Educational and Behavioral Statistics. 2013;38(5):499 521.

[30] G. J. McLachlan and D. Peel. Finite mixture models. John Wiley & Sons; 2000.

[31] D. Vidotto, J. K. Vermunt and M.C. Kaptein. Multiple imputation of missing categorical data using latent class models: state of the art. Psychological test and assessment modeling. 2015;57(4):542.

[32] D. W. van der Palm DW, L. A. van der Ark, J. K. Vermunt. A comparison of incomplete data methods for categorical data. Statistical Methods in Medical Research. 2016;25(2):754 774. PMID: 23166159. Available from: https://doi.org/10.1177/0962280212465502.

[33] M. Sadinle. Detecting duplicates in a homicide registry using a Bayesian partitioning approach. The Annals of Applied Statistics. 2014;8(4):2404 2434.

[34] M. Sadinle. Bayesian estimation of bipartite matchings for record linkage. Journal of the American Statistical Association. 2017;112(518):600 612.

[35] M. Sadinle. Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations. The Annals of Applied Statistics. 2018;12(2):1013 1038.

[36] R. C. Steorts, Hall R, Fienberg SE. A Bayesian approach to graphical record linkage and deduplication. Journal of the American Statistical Association. 2016;111(516):1660 1672.

[37] R. C. Steorts, Entity resolution with empirically motivated priors. Bayesian Analysis. 2015;10(4):849 875.

[38] D. Manrique-Vallier and J. P. Reiter. Bayesian estimation of discrete multivariate latent structure models with structural zeros. Journal of Computational and Graphical Statistics. 2014;23(4):1061 1079.

[39] P. Lazarsfeld. The logical and mathematical foundation of latent structure analysis. Studies in Social Psychology in World War II Vol IV: Measurement and Prediction. 1950:362 412.

[40] L. A. Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika. 1974 08;61(2):215 231. Available from: https://doi.org/10.1093/biomet/61.2.215.

[41] M. Gebregziabher and S. M. DeSantis. Latent class based multiple imputation approach for missing categorical data. Journal of Statistical Planning and Inference. 2010;140(11):3252 3262.

[42] D. Vidotto, J. Vermunt and K. Van Deun. Bayesian latent class models for the multiple imputation of categorical data. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences. 2018;14(2):56 68.

[43] D. W. Van der Palm, L. A. Van der Ark and J. K. Vermunt. Divisive latent class modeling as a density estimation method for categorical data. Journal of Classification. 2016;33(1):52 72.

[44] D. Vidotto, J. Vermunt and K. van Deun. Bayesian multilevel latent class models for the multiple imputation of nested categorical data. Journal of Educational and Behavioral Statistics. 2018;43(5):511 539.

[45] D. B. Rubin. Multiple imputation for nonresponse in surveys. vol. 81. John Wiley & Sons; 1987.

[46] J. Sethuraman. A constructive definition of Dirichlet priors. Statistica sinica. 1994:639 650.

[47] H. Ishwaran and L.F. James. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association. 2001;96(453):161 173.

[48] D. B. Dunson and C. Xing. Nonparametric Bayes modeling of multivariate categorical data. Journal of the American Statistical Association. 2009;104(487):1042 1051.

[49] M. D'Orazio. Package StatMatch. Statistical matching or data fusion, version. 2019;1(0).

[50] I. Lewaa, M. S. Hafez and M. A. Ismail. Statistical Matching of HICES and EDHS: A Mixed Approach. Mathematics and Statistics. 2023.

Appendix A:

Table 1: Missing Pattern of SM.

	X	Y	Z
File A	data	data	?
File B	data	?	data

Table 2: Probabilities for each category of X variables.

	Category 1	Category 2	Category 3	Category 4
X_1	0.3	0.3	0.3	0.1
X_2	0.3	0.3	0.4	-
X_3	0.4	0.6	-	-

Table 3: Cramer's V with Y under CIA.

Variables	Cramer's V with Y	Variables	Cramer's V with Y
Z_1	0.005163	Z_6	0.005835
Z_2	0.04815	Z_7	0.04769
Z_3	0.08354	Z_8	0.07503
Z_4	0.1356	Z_9	0.1335
Z_5	0.1812	$Z10$	0.1781

Table 4: Coefficients β'_h and λ'_h in 5.3.

	h	β'_h coefficients of x's	λ'_h coefficients of y's
Z_1	1 2 3	(0,0,0,0,0,0,0,0,0) (0,0,0,0,0,0,0,0,0) (0,0,0,0,0,0,0,0,0)	(0,0,0) (0,0,0) (0,0,0)
Z_2	1 2 3	(0,0,0,0,0,0,0,0,0) (0,0,0,0,0,0,0,0,0) (0,0,0,0,0,0,0,0,0)	(0.3,0,0.3) (0,0.3,0) (0,0.3,0)

Z_3	1 2 3	(0,0,0,0,0,0,0,0,0,0) (0,0,0,0,0,0,0,0,0,0) (0,0,0,0,0,0,0,0,0,0)	(0.5,0,0.5) (0,0.5,0) (0,0.5,0)
Z_4	1 2 3	(0,0,0,0,0,0,0,0,0,0) (0,0,0,0,0,0,0,0,0,0) (0,0,0,0,0,0,0,0,0,0)	(0.8,0,0.8) (0,0.8,0) (0,0.8,0)
Z_5	1 2 3	(0,0,0,0,0,0,0,0,0,0) (0,0,0,0,0,0,0,0,0,0) (0,0,0,0,0,0,0,0,0,0)	(1,0,1) (0,1,0) (0,1,0)
Z_6	1 2 3	(0.2,0,0.2,0,0.2,0,0.2,0,0.2) (0,−0.2,0,−0.2,0,−0.2,0,−0.2,0) (0.2,0,0.2,0,0.2,0,0.2,0,0.2)	(0,0,0) (0,0,0) (0,0,0)
Z_7	1 2 3	(0.2,0,0.2,0,0.2,0,0.2,0,0.2) (0,−0.2,0,−0.2,0,−0.2,0,−0.2,0) (0.2,0,0.2,0,0.2,0,0.2,0,0.2)	(0.3,0,0.3) (0,0.3,0) (0,0.3,0)
Z_8	1 2 3	(0.2,0,0.2,0,0.2,0,0.2,0,0.2) (0,−0.2,0,−0.2,0,−0.2,0,−0.2,0) (0.2,0,0.2,0,0.2,0,0.2,0,0.2)	(0.5,0,0.5) (0,0.5,0) (0,0.5,0)
Z_9	1 2 3	(0.2,0,0.2,0,0.2,0,0.2,0,0.2) (0,−0.2,0,−0.2,0,−0.2,0,−0.2,0) (0.2,0,0.2,0,0.2,0,0.2,0,0.2)	(0.8,0,0.8) (0,0.8,0) (0,0.8,0)
$Z10$	1 2 3	(0.2,0,0.2,0,0.2,0,0.2,0,0.2) (0,−0.2,0,−0.2,0,−0.2,0,−0.2,0) (0.2,0,0.2,0,0.2,0,0.2,0,0.2)	(1,0,1) (0,1,0) (0,1,0)

Table 5: Simulation Results for SM Under CIA.

	Loglinear	MICE	Random Hotdeck	KP1	KP2	MDC1	MDC2	MDC3	MVR
Z1	0.067916 0.082277 0.043554	0.63703 0.671413 0.56279	0.046298 0.5989861 0.03269765	0.03728109 0.04499786 0.03056432	0.03849114 0.04555581 0.03142646	0.03785468 0.04108886 0.02686671	0.04079668 0.04472665 0.02894818	0.04203552 0.04834557 0.02572548	0.0499349 0.053764 0.038223
Z2	0.077717 0.099238 0.056196	0.635212 0.658184 0.566535	0.04369243 0.06617226 0.02121259	0.0314322 0.03900372 0.02886069	0.02930849 0.04259478 0.02602221	0.0352827 0.04119779 0.02936762	0.077717 0.099238 0.056196	0.635212 0.658184 0.566535	0.04369243 0.06617226 0.02121259
Z3	0.11036802 0.13584062 0.08489542	0.630323 0.676064 0.492603	0.05084272 0.08281083 0.04187461	0.05465757 0.07046931 0.04884582	0.05536806 0.07614948 0.04458664	0.03958162 0.04588188 0.03428136	0.0412471 0.05234978 0.03014441	0.0426692 0.05883201 0.03870183	0.06012001 0.0810091 0.04613911
Z4	0.08131489 0.16040252 0.04222727	0.61790 0.693001 0.534036	0.0819831 0.1332374 0.064247	0.06966446 0.07702479 0.06230413	0.06811114 0.07752968 0.0586926	0.05188 0.054162 0.048136	0.052471 0.0584978 0.0414441	0.0426692 0.04883201 0.040183	0.0823043 0.14320842 0.07325244

Z5	0.12077104 0.15098475 0.09055734	0.67794 0.709662 0.632049	0.10021642 0.18101031 0.08942254	0.0992836 0.19242963 0.08613756	0.09908646 0.17781276 0.0836015	0.0602015 0.067453 0.05456577	0.06006326 0.06508155 0.0554496	0.060843 0.065134 0.05718068	0.0837041 0.16328443 0.07345638
Z6	0.07289756 0.11490919 0.04088594	0.569763 0.66106 0.314710	0.03516833 0.05172303 0.02861364	0.02755114 0.03085608 0.01424619	0.02620426 0.03075644 0.01165209	0.03916137 0.04102168 0.02730106	0.04414645 0.06467827 0.02361462	0.03881842 0.05442842 0.03320842	0.05684898 0.0717183 0.04652613
Z7	0.08526856 0.11024027 0.05029685	0.615340 0.668350 0.506462	0.04663235 0.06205305 0.03121164	0.03151987 0.04573381 0.02730594	0.03277866 0.04073465 0.02482267	0.04168211 0.06632415 0.03704007	0.03551746 0.04318987 0.02784506	0.04934159 0.05240422 0.02627895	0.04279541 0.04552737 0.04006346
Z8	0.09463486 0.12712072 0.052149	0.595460 0.648939 0.425084	0.07553193 0.08685231 0.0621156	0.0552034 0.07022461 0.05018219	0.063564 0.0756431 0.0502557	0.05194293 0.05528485 0.04860102	0.04458001 0.05069987 0.03846015	0.04835613 0.05882636 0.03788589	0.05912837 0.07118472 0.03707203
Z9	0.12489024 0.14602452 0.0875595	0.602766 0.69391 0.15299	0.083206 0.121167 0.0694295	0.08457138 0.0932902 0.06481373	0.08665002 0.09405619 0.07924385	0.0590296 0.0611483 0.05703109	0.06377618 0.0703394 0.05991842	0.06075612 0.070992 0.0541304	0.08391172 0.0975727 0.06025073
Z10	0.12041956 0.19095544 0.07988368	0.586799 0.710990 0.308525	0.0869765 0.11145372 0.07249929	0.09682839 0.18794658 0.0857102	0.09480808 0.19174123 0.07787492	0.08172354 0.09456929 0.06887778	0.07462893 0.08261829 0.07063957	0.07266273 0.0816315 0.06369395	0.093989 0.185872 0.081776

Mean of Hellinger distance is the first value in each cell. Max is the second value in each cell and min is the third value in each cell.

Table 6: Simulation Results for SM with Auxiliary Information.

	Loglinear	MICE	Random hotdeck	KP1	KP2	MDA1	MDA2	MDA3	MVR
Z1	0.05860018 0.15651979 0.04068057	0.21849771 0.42435584 0.11263958	0.08678879 0.09678879 0.0788536	0.07497618 0.09255556 0.0473968	0.04707412 0.07252511 0.02762313	0.03093397 0.04512779 0.01988734	0.26432687 0.30256828 0.12608546	0.03882768 0.0539806 0.02252758	0.081709 0.184082 0.069934
Z2	0.07294898 0.14665598 0.05924198	0.17046792 0.34211482 0.0882102	0.07294898 0.14665598 0.05924198	0.08517152 0.13830146 0.05204158	0.0539654 0.08735677 0.03057403	0.03732575 0.0507664 0.0208851	0.14342608 0.27527641 0.08157576	0.04838788 0.06379936 0.03297639	0.170433 0.274067 0.09468
Z3	0.07774732 0.18115698 0.05433765	0.271485 0.327849 0.11512181	0.089274 0.181642 0.061642	0.08617857 0.1126907 0.06383794	0.07136818 0.09188031 0.05085604	0.04372872 0.05890749 0.02754995	0.15133253 0.17492825 0.12773681	0.0534728 0.07215524 0.04153931	0.13603469 0.196704 0.08536468
Z4	0.07201562 0.09082011 0.04321114	0.20639037 0.30864942 0.10413132	0.0962756 0.199735 0.083161	0.07243384 0.18935663 0.04551105	0.08206373 0.09424444 0.05988303	0.04603707 0.0563539 0.03572023	0.2959781 0.47095177 0.12100443	0.0594377 0.07947586 0.04241168	0.230031 0.32441 0.14762
Z5	0.07496258 0.08482725 0.03509791	0.2921363 0.39798472 0.18628788	0.081572 0.1881567 0.062518	0.06633839 0.07992855 0.05274823	0.09539153 0.19329497 0.0648808	0.03679873 0.05102797 0.03080034	0.21363374 0.4884706 0.17842041	0.0552615 0.07565079 0.0348722	0.221093 0.36192 0.115994

Z6	0.07333214 0.1440407 0.05262357	0.2713763 0.38672354 0.09560290	0.083566 0.095543 0.070431	0.08229237 0.09164205 0.0629427	0.05080219 0.09710279 0.0445016	0.03338099 0.04108732 0.02567465	0.20381926 0.38679704 0.12084149	0.04417221 0.05330034 0.03904407	0.18145111 0.254045 0.09536176
Z7	0.0843678 0.0953985 0.06919576	0.25810676 0.39634087 0.11987266	0.0711378 0.091367 0.061962	0.06501911 0.1256244 0.04251382	0.05690486 0.0717694 0.03204033	0.0486462 0.05971151 0.02758089	0.29778225 0.36194741 0.13361709	0.05417467 0.07614986 0.03219948	0.18420716 0.2215117 0.13126316
Z8	0.07333214 0.1440407 0.05262357	0.2713763 0.38672354 0.09560290	0.083566 0.095543 0.070431	0.08229237 0.09164205 0.0629427	0.05080219 0.09710279 0.0445016	0.03338099 0.04108732 0.02567465	0.20381926 0.38679704 0.12084149	0.04417221 0.05330034 0.03904407	0.18145111 0.254045 0.09536176
Z9	0.0843678 0.0953985 0.06919576	0.25810676 0.39634087 0.11987266	0.0711378 0.091367 0.061962	0.06501911 0.1256244 0.04251382	0.05690486 0.0717694 0.03204033	0.0486462 0.05971151 0.02758089	0.29778225 0.36194741 0.13361709	0.05417467 0.07614986 0.03219948	0.18420716 0.2215117 0.13126316
Z10	0.06993457 0.0979791 0.04189004	0.14871528 0.26682716 0.08060339	0.084336 0.153453 0.075228	0.08373355 0.16495026 0.05551684	0.07418653 0.09708243 0.05129064	0.04477651 0.05701776 0.03253526	0.22176578 0.35553861 0.18799296	0.0598455 0.07004128 0.04192781	0.182517 0.238547 0.11184

Mean of Hellinger distances is the first value in each cell. Max is the second value in each cell and min is the third value in each cell.

Table 7: Simulation Results for Different Sizes of Auxiliary File.

	MDA1 (50)	MDA1 (100)	MDA1 (150)	MDA1 (200)	MDA1 (250)	MDA1 (300)
Z1	0.098109368 0.160126571 0.073341718	0.071061779 0.131270529 0.04060588	0.051248922 0.077334062 0.027285611	0.043278719 0.060839894 0.02982578	0.037352552 0.054399649 0.0257082	0.037030207 0.04802272 0.022063264
Z2	0.103272073 0.197755475 0.05625384	0.06799996 0.091258791 0.048993396	0.059121331 0.075890099 0.038639812	0.053013808 0.065143537 0.044968045	0.050141797 0.06148519 0.041850571	0.047689362 0.07206718 0.034132988
Z3	0.094458412 0.144034225 0.072102415	0.078307309 0.107005078 0.039411749	0.060703085 0.072439343 0.038704251	0.054983171 0.065081818 0.043513728	0.04934298 0.063396294 0.036986699	0.04526447 0.063167138 0.034396257
Z4	0.133150681 0.227618346 0.082778178	0.085808109 0.114923109 0.055122075	0.070462961 0.085938214 0.058602386	0.064636378 0.082877845 0.050550607	0.067026812 0.095525106 0.043829276	0.066590987 0.086764917 0.047186441
Z5	0.11906136 0.142469812 0.073604721	0.090863698 0.148703532 0.060961913	0.079767331 0.104160207 0.04252481	0.077711 0.111021 0.051708	0.075800624 0.098870823 0.039853513	0.070156669 0.091510174 0.033731452

Z6	0.097600913 0.147806324 0.063088474	0.077459382 0.102661443 0.047053785	0.057169133 0.074209599 0.038019592	0.055738205 0.083230584 0.037980021	0.055096824 0.076324005 0.03721917	0.047885603 0.069021589 0.025097529
Z7	0.100043673 0.219532314 0.054407544	0.081219987 0.136969003 0.047275786	0.069940363 0.10613925 0.038348415	0.058548493 0.085921233 0.038924989	0.057921688 0.075299801 0.038168056	0.049499607 0.069985869 0.027358598
Z8	0.121700648 0.164643972 0.07622247	0.081901793 0.131927022 0.031632467	0.070446854 0.091915559 0.043748742	0.056288776 0.078801894 0.03820131	0.053665943 0.080037957 0.037915987	0.050043616 0.083172389 0.036208435
Z9	0.103835753 0.134886387 0.067249566	0.077011004 0.101920568 0.029219054	0.076564874 0.094090749 0.053889218	0.073217681 0.083289272 0.055793291	0.06799713 0.085232831 0.05383064	0.066710014 0.081322874 0.049416539
Z10	0.091240012 0.146050281 0.052897858	0.079894186 0.095169789 0.052502195	0.080329596 0.109051814 0.044910685	0.073708494 0.104277657 0.034529324	0.07051435 0.101543175 0.035368397	0.069377981 0.096908817 0.040682222

Mean of Hellinger distance is the first value in each cell. Max is the second value in each cell and min is the third value in each cell.

Appendix B:

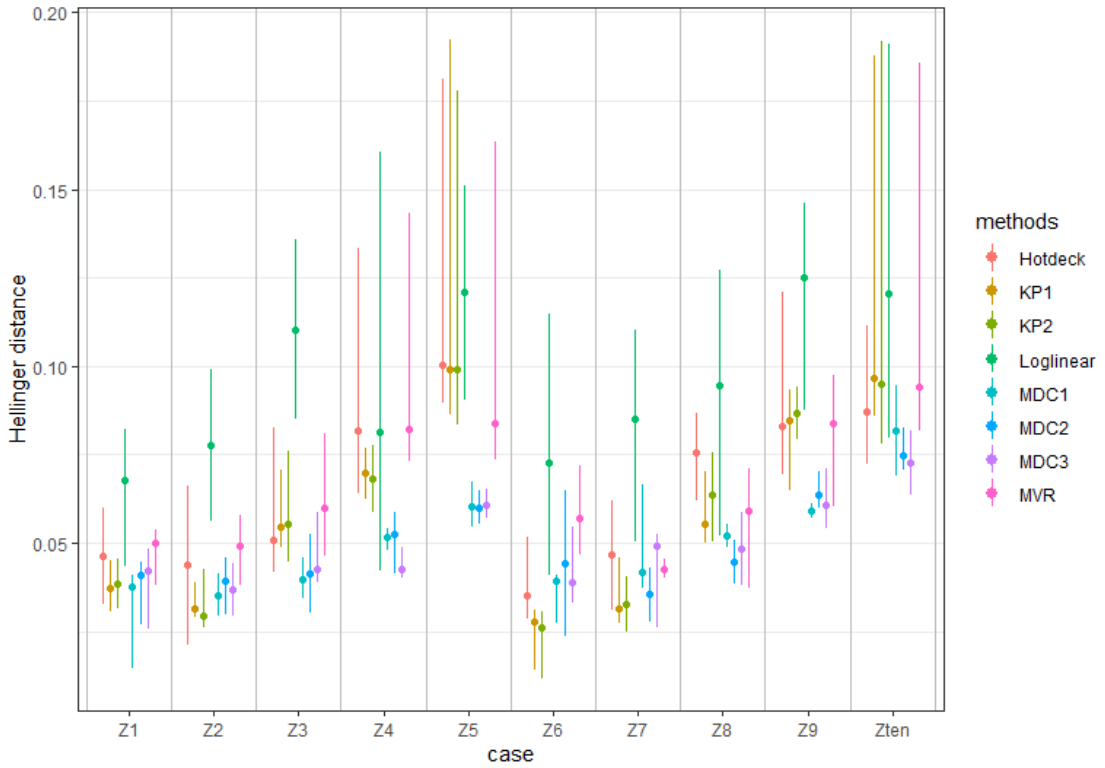

Fig.1: Comparison between different approaches under CIA.

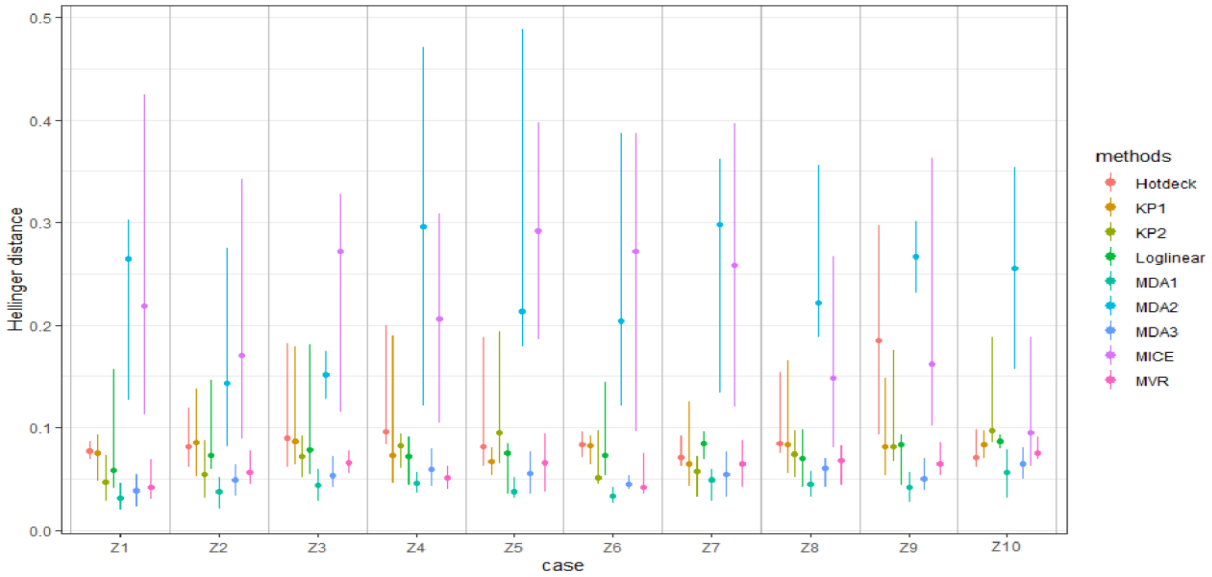

Fig. 2: Comparison between different approaches with Auxiliary Information.

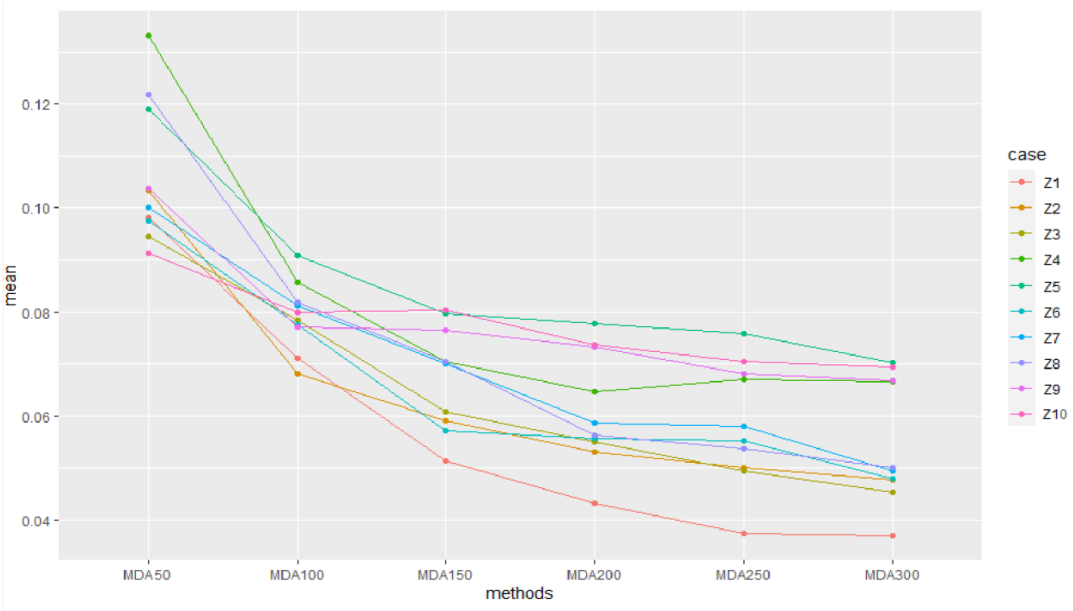

Fig. 3: Comparison of Hellinger distance between MDA1 with various sample size of file.