Psychological Test and Assessment Modeling, Volume 57, 2015 (4), 577-594

Partitioned predictive mean matching as a multilevel imputation technique

Gerko Vink[1], *Goran Lazendic*[2] *& Stef van Buuren*[3]

Abstract

Large scale assessment data often has a multilevel structure. When dealing with missing values, such structures need to be taken into account to prevent underestimation of the intraclass correlation. We evaluate predictive mean matching (PMM) as a multilevel imputation technique and compare it to other imputation approaches for multilevel data. We propose partitioned predictive mean matching (PPMM) as an extension to the PMM algorithm to divide the big data multilevel problem into manageable parts that can be solved by standard predictive mean matching. We show that PPMM can be a very effective imputation approach for large multilevel datasets and that both PPMM and PMM yield plausible inference for continuous, ordered categorical, or even dichotomous multilevel data. We conclude that both the performance of PMM and PPMM is often comparable to dedicated methods for multilevel data.

Keywords: Large datasets, Multilevel data, Multiple imputation, Partitioning, Predictive mean matching.

[1] *Correspondence concerning this article should be addressed to:* dr. Gerko Vink, Department of Methods and Statistics, Utrecht University, Padualaan 14, 3584CH Utrecht, the Netherlands; email: G.Vink@uu.nl

[2] Australian Curriculum, Assessment and Reporting Authority

[3] Utrecht University and Netherlands Organization for Applied Scientific Research TNO

Introduction

In large scale assessment surveys, missing values for student demographic and socioeconomic background data are frequently encountered. Often such data has a multilevel structure, where respondents are nested within naturally occurring clusters, such as schools or municipalities. Accounting for missingness in data with a multilevel structure is a relatively recent development, and much remains unknown.

In the multilevel analysis model, cluster effects are assumed to be random. Simply ignoring the effects of the cluster during imputation can lead to an underestimation of the intraclass correlation (ICC) in the completed data. Ideally, model parameters within the clusters are allowed to randomly vary during imputation. Although such strategies are available (Zhao and Schafer, 2013), they rely heavily on model assumptions.

Some authors have indicated that multilevel data can also be reasonably imputed when cluster membership is taken into account as a fixed effect (Andridge, 2011; Graham, 2012). A straightforward procedure to include cluster membership into the imputation model as a fixed effect is to use dummy coding strategies. Such strategies should allow for proper estimation of the intraclass correlation, especially when the ICC gets large (Andridge, 2011; Graham, 2012).

Imputing multilevel data by incorporating a fixed effects approach in the imputation model can be very convenient in practice. In such situations, a straightforward extension of current imputation methodology to the situation of multilevel data suffices. For real life data, which does not necessarily follow a specific distribution, such an extension can be especially flexible when an approach is used that does not pose strict assumptions on the distribution of the data.

One imputation approach that is particularly proven to work well in a wide range of situations is predictive mean matching (PMM, Little, 1988; Rubin, 1986). It has been shown that the performance of imputation procedures involving PMM can be very good (De Waal, Pannekoek and Scholtus, 2011; Siddique and Belin, 2007; Su, Gelman, Hill and Yajima, 2011; Van Buuren, 2012; Van Buuren and Groothuis-Oudshoorn, 2011; White, Royston and Wood, 2011; Yu, Burton and Rivero-Arias, 2007), especially when normality assumptions are breached or when semicontinuous data is considered (Van Buuren, 2012; Vink, Frank, Pannekoek and Buuren, 2014). More specifically, PMM does not only yield acceptable and plausible estimates, but also manages to maintain underlying distributions of the data (Heeringa, Little and Raghunathan, 2002; Van Buuren, 2012; Vink, Frank, Pannekoek and Buuren, 2014; White, Royston and Wood, 2011; Yu, Burton and Rivero-Arias, 2007). Implementation of PMM is straightforward in multivariate data problems when the fully conditional specification (FCS, Van Buuren,

Brand, Groothuis-Oudshoorn and Rubin, 2006) framework is used. Little is known, however, about the practical applicability of PMM on multilevel data.

We investigate how suitable PMM is as an imputation approach for multilevel data when the clustering of the data is modeled as a fixed effect during imputation. We thereby concentrate on a comparison between PMM, bayesian normal linear imputation (NORM) and a mixture of suitable, dedicated 2-level imputation methods (MIX) in a simulation study. We propose partitioned predictive mean matching (PPMM) as an extension to the predictive mean matching algorithm to facilitate imputation of multilevel data for big datasets. Finally, we apply PPMM to a large real dataset from the Australian Curriculum, Assessment and Reporting Authority (ACARA) to obtain a Bayesian estimate of Social-Educational Advantage on the school level.

Predictive mean matching as a multilevel imputation approach

We define $Y = (Y_{obs}, Y_{mis})$ as an incomplete variable, where Y_{obs} and Y_{mis} denote the observed values and the missing values in Y, respectively. We define $X = (X_1, ..., X_j)$ as a set of j fully observed covariates, with X_{obs} and X_{mis} corresponding to the observed and missing parts in Y. Further, n denotes the number of units in Y, n_{mis} and n_{obs} denote the number of units with missing and observed values of Y, respectively, and m denotes the number of multiply imputed datasets to be obtained, with $m \geq 2$. Finally, we require the variable that contains the class structure to be included in X in dummy coded form.

PMM algorithm

Multiply imputing Y_{mis} by means of predictive mean matching can be done by the following algorithm.

1. Use linear regression of Y_{obs} given X_{obs} to estimate $\hat{\beta}$, $\hat{\sigma}$ and $\hat{\varepsilon}$ by means of ordinary least squares.
2. Draw σ^{2*} as $\sigma^{2*} = \hat{\varepsilon}^T \hat{\varepsilon}/A$, where A is a χ^2 variate with $n_{obs} - j$ degrees of freedom.
3. Draw β^* from a multivariate normal distribution centered at $\hat{\beta}$ with covariance matrix $\sigma^{2*}(X_{obs}^T X_{obs})^{-1}$.
4. Calculate $\hat{Y}_{obs} = X_{obs}\,\hat{\beta}$ and $\hat{Y}_{mis} = X_{mis}\,\beta^*$.
5. For each missing value $\hat{Y}_{mis,i}$ where $i = 1, \ldots, n_{mis}$:
   a) find $\Delta = |\hat{Y}_{obs,k} - \hat{Y}_{mis,i}|$ for all k, with $k = 1, \ldots, n_{obs}$.

b) Randomly sample one value from $(\Delta^{(1)}, \ldots, \Delta^{(5)})$, where $\Delta^{(1)}$ through $\Delta^{(5)}$ are the five smallest elements in Δ, respectively, and take the corresponding Y_{obs} as the imputation.

6. Repeat Steps 1 through 5 m times, each time saving the completed dataset.

In the case of multivariate missingness, FCS can be used to iteratively impute every missing datum in each variable of interest, based on a set of covariates. We must note that alternative implementations of PMM do exist (see e.g. Koller-Meinfelder (2009); Morris, White and Royston (2014); Schenker and Taylor (1996); Siddique and Belin (2007)).

Selecting donors

When performing PMM on multilevel data, three possible scenarios can be used to sample a probable donor value. First, if we ignore the cluster structure, any value in Y_{obs} can in theory be sampled as a donor value, although some values are more likely than others. Assumed that units within a cluster are more alike than units between clusters, this scenario will ignore valuable information, potentially leading to biased results and underestimated cluster effects.

Alternatively, missing values in Y can be imputed by sampling a suitable donor candidate from within the respective cluster. Although potentially very effective, this scenario will quickly pose donor selection problems when clusters size becomes too small or when clusters are completely unobserved.

We prefer a compromise between the first two scenarios. In the algorithm from , the cluster structure is included in the prediction models for $\hat{Y}_{mis}$ and $\hat{Y}_{obs}$. As a result, the likelihood of a sampled donor value coming from the same cluster (or a similar cluster, for that matter) as the missing value is increased. In this way, the cluster structure is preserved as far as possible, while still allowing for probable donor selection in the case of very small or completely unobserved clusters.

Partitioned predictive mean matching (PPMM)

As datasets become increasingly larger, using dummy coding strategies can become computationally challenging. In the case of many respondents (say 3 million) combined with a large number of clusters (say 10 thousand), expanding the cluster structure to

dummy variables may currently even be computationally unfeasible. To avoid computational problems when using predictive mean matching with large multilevel datasets, we propose the following extension to the predictive mean matching algorithm:

1. Partition the data into P approximately equally sized smaller parts, where each part $p = 1, \ldots, P$ contains only whole clusters.

2. Carry out the PMM algorithm from page 579 on each part p.

3. Append the P parts for each multiply imputed dataset.

Without loss of generality, the combined data for the P imputed parts can be analyzed conform to current imputation methodology. For the estimation process it is critical that clusters are wholly contained in a single part and are not split among parts. If the data is ordered based on a set of (observed) covariates, such that the likelihood of similar clusters being in the same part is increased, selecting a probable donor can be done on the level of the available donors, without the need for the data as a whole. For example, demographic information can be used to group similar clusters into the same parts. Such a procedure would benefit the imputation model, especially when donor candidates need to be sampled from outside the 'own' cluster.

Speeding up donor selection

Both PMM and PPMM draw imputations from observed values by comparing the distance between each $\hat{Y}_{mis}$ with all $\hat{Y}_{obs}$. This process can become a very lengthy procedure for very large datasets (e.g. $n > 1,000,000$). Using a sufficiently large randomly selected subsample from $\hat{Y}_{obs}$ to sample donors from is computationally convenient and efficient, especially when the number of cases and the proportion of missingness are both large. We propose the following extension to the donor selection step in the PMM algorithm from page 579 for large datasets with large amounts of missingness:

1. Draw a subsample $\hat{Y}^S_{obs}$ of length l randomly from $\hat{Y}_{obs}$, with $l < n_{obs}$.

2. Find for each missing value $\hat{Y}_{mis,i}$ the five smallest donors from $\Delta = |\hat{Y}^S_{obs,k_1} - \hat{Y}_{mis,i}|$, where $k_1 = 1, \ldots, l$.

We must note that it is not necessary to choose $l < n_{obs}$, but doing so may greatly benefit computation time.

Simulation

Predictive mean matching is an imputation method that is relatively easy to implement with a performance that is often very good. To gain insight in the suitability of PMM and PPMM as a multiple imputation approach for multilevel data we performed the following simulation study.

We use the `popularity2` dataset from Hox (2010), a simulated dataset for 2,000 pupils in 100 classes. The dataset contains two level one outcome measures that consider pupil popularity; an indication of pupil popularity (popular, $\mu = 5.08$) derived by a sociometric procedure and pupil popularity as perceived by the teacher (popteach, $\mu = 5.06$). Both outcome variables are measured on a 10-point scale. The explanatory variables are pupil gender (sex, $\mu = 0.51$), pupil's self-measured extraversion (extrav, $\mu = 5.22$) on a 10-point scale, and the experience of the teacher (texp, $\mu = 14.26$) measured in years. The `popularity` data does not consider the school level.

We induce missing completely at random (MCAR) and missing at random (MAR) missingness in the popularity data based on popularity as perceived by the teacher (popteach). Missing values are assigned by using a random draw from a binomial distribution of the same length as Y and of size 1 following the procedure as described in Vink et al. (2014). In the simulations, 15 %, 25 % and 50 % missingness is induced.

To simulate PPMM we partition the data into 10 parts, where each part contains roughly 200 pupils and classes are not split across parts. The average self-perceived pupil popularity differs greatly across classes, which may result in poor performance of the imputation method when using random partitioning, especially when the amount of missingness is large. To this end, the data are sorted based on the average pupil popularity in each class, such that average pupil popularity within parts is more similar than average pupil popularity between parts.

Data imputations are performed with `mice` (version 2.21, Van Buuren and Groothuis-Oudshoorn, 2011) in `R` (version 3.1.0, R Core Team, 2013), with 5 multiply imputed datasets and 10 iterations for the algorithm to converge. PMM and PPMM (both performed by `mice.impute.fastpmm`) are compared to a distributional approach (normal bayesian linear imputation conform `mice.impute.norm`) and a mix of dedicated multilevel imputation algorithms (MIX), namely `mice.impute.2l.norm` for 'extrav', `mice.impute.2l.pan` (based on PAN by Zhao and Schafer (2013)) for 'popular' and `mice.impute.2lonly.mean` for 'texp'. We use logistic regression imputation (`mice.impute.logreg`) to impute 'gender' for NORM and MIX, but leave out the results under MIX as they are equivalent to the results under NORM.

Table 1:
Overview of imputation methods used per variable in the simulation

	LEVEL 1			LEVEL 2
	extrav	sex	popular	texp
PMM	pmm	pmm	pmm	pmm
NORM	norm	logreg	norm	norm
MIX	2l.norm	-	2l.pan	2lonly.mean

Each cluster in the popularity dataset contains at least 16 and at most 26 students, with the mode being 20 students. There is no need to investigate larger cluster sizes, as it is known that bias in the ICC decreases as cluster size gets larger (Andridge, 2011).

Evaluation

We evaluate the imputation approaches on the ability to retrieve the following model components for each variable: the average bias of the group means (fixed effect bias), the average coverage rate of the 95 % confidence interval of the group means and the ICC. The ICC is defined as $\sigma_\alpha^2/(\sigma_\alpha^2+\sigma_\varepsilon^2)$ where σ_α^2 denotes the between group variance and σ_ε^2 denotes the within group variance of the random effects. Conveniently, the ICC contains the information about the random effects variance components, therefor we do not evaluate these variances separately.

The above evaluations are carried out on the variable level instead of the model level for two reasons. First, we would like to preserve data structures. Second, the role variables take in a model might change during the analysis stage. Outcome variables in one model may become predictors in another model, and vice versa. Ultimately, it would be ideal if both models can be analyzed on the same data.

Because we consider the popularity data as the population and induce missingness in the data directly, we have no sampling variation and the pooling rules proposed in Vink and Van Buuren (2014) are used.

Results

Intraclass correlation

The popularity data displays strong population intraclass correlations (see Table 2). All imputation methods yield results that are relating close to these population values. As expected, the bias increases with the amount of missingness.

The experience of the teacher (texp) is particularly interesting when considering the intraclass correlation. Because the experience is the same for all pupils in a cluster, the intraclass correlation equals 1. PMM is able to automatically replicate this structure, yielding the correct inference as if the data were deductively imputed. NORM does not sample from the observed data, but rather draws from a normal distribution, resulting in a small deviation from the population value. MIX uses a cluster mean imputation method, which yields unbiased results when at least one pupil is observed in each cluster. For larger amounts of missingness, however, it may occur that clusters are completely unobserved. In such situations, MIX is not able to find an imputation.

Table 2:
Bias of the intraclass correlations after imputation as deviations from the population value (truth).

		15 % missing				25 % missing				50 % missing			
meth	mech	extrav	sex	texp	pop	extrav	sex	texp	pop	extrav	sex	texp	pop
TRUTH	-	0.262	0.112	1	0.363	0.262	0.112	1	0.363	0.262	0.112	1	0.363
PMM	mcar	0.009	0.002	0	0.004	0.017	0.009	0	0.008	0.047	0.036	0	0.021
	left	0.007	0.004	0	0.002	0.014	0.010	0	0.004	0.053	0.048	0	0.017
	right	0.011	0.002	0	0.001	0.020	0.008	0	0.006	0.056	0.044	0	0.020
	mid	0.009	0.003	0	0.004	0.017	0.010	0	0.008	0.050	0.046	0	0.025
	tail	0.009	0.002	0	0.000	0.015	0.006	0	0.003	0.050	0.031	0	0.015
PPMM	mcar	0.007	0.001	-0.001	0.006	0.012	0.004	-0.001	0.011	0.034	0.027	-0.003	0.030
	left	0.006	0.005	-0.001	0.005	0.016	0.013	-0.001	0.007	0.048	0.047	-0.002	0.029
	right	0.004	-0.009	-0.001	0.006	0.013	-0.012	-0.002	0.009	0.033	0.011	-0.007	0.030
	mid	0.007	0.004	-0.001	0.004	0.012	0.011	-0.001	0.008	0.035	0.041	-0.002	0.031
	tail	0.005	-0.004	-0.001	0.004	0.010	-0.004	-0.001	0.007	0.034	0.011	-0.003	0.026
NORM	mcar	0.007	-0.009	≈ 0	0.003	0.012	-0.009	≈ 0	0.004	0.043	0.013	≈ 0	0.018
	left	0.006	-0.009	≈ 0	0.001	0.011	-0.004	≈ 0	0.003	0.042	0.036	≈ 0	0.015
	right	0.011	-0.011	≈ 0	0.001	0.020	-0.010	≈ 0	0.005	0.057	0.024	≈ 0	0.022
	mid	0.007	-0.003	≈ 0	0.003	0.014	-0.001	≈ 0	0.006	0.042	0.026	≈ 0	0.018
	tail	0.007	-0.016	≈ 0	-0.000	0.014	-0.017	≈ 0	0.001	0.048	0.007	≈ 0	0.011
MIX	mcar	-0.001	-	0	-0.000	-0.006	-	0	-0.001	-0.018	-	0	0.003
	left	-0.006	-	0	-0.002	-0.012	-	0	-0.003	-0.030	-	*0	0.001
	right	-0.006	-	0	-0.000	-0.011	-	0	0.002	-0.019	-	*0	0.011
	mid	-0.001	-	0	-0.000	-0.005	-	0	-0.000	-0.020	-	0	0.005
	tail	-0.007	-	0	-0.000	-0.012	-	0	0.000	-0.031	-	0	0.003

*Values are calculated based on the imputed clusters only, due to some clusters being completely unobserved.

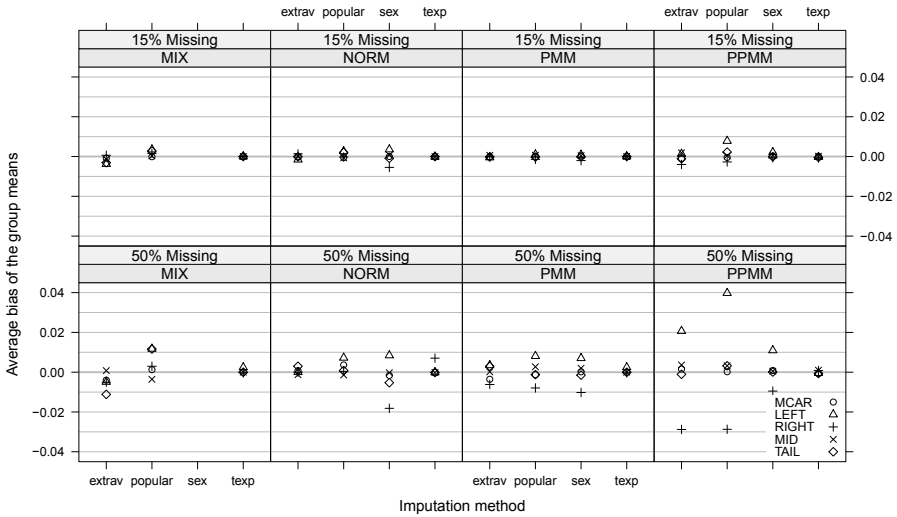

Figure 1:
Average bias of the group means. Shown are results for four imputation approaches and four variables for varying missingness percentages.

The slightly larger bias for the ICC for teacher experience in the case of PPMM can be explained by the correlation between pupil popularity and teacher experience ($\rho(1998) = .29, p < .01$). Sorting the data based on pupil popularity will have an effect on the distribution of teacher experience over parts. Together with the smaller sample size, this results in occasional imputations for teacher experience that are different from the observed values.

Fixed effects

The fixed effects are very accurately estimated by all imputation approaches, see Figure 1. PMM displays a very stable performance, with very low variation in the bias across missingness mechanisms. PPMM shows more variation in the case of 50 % missing data, which can be explained by restrictions put ont the available donor pool due to partitioning. For all methods it holds that bias is very small, even for large amounts of missingness.

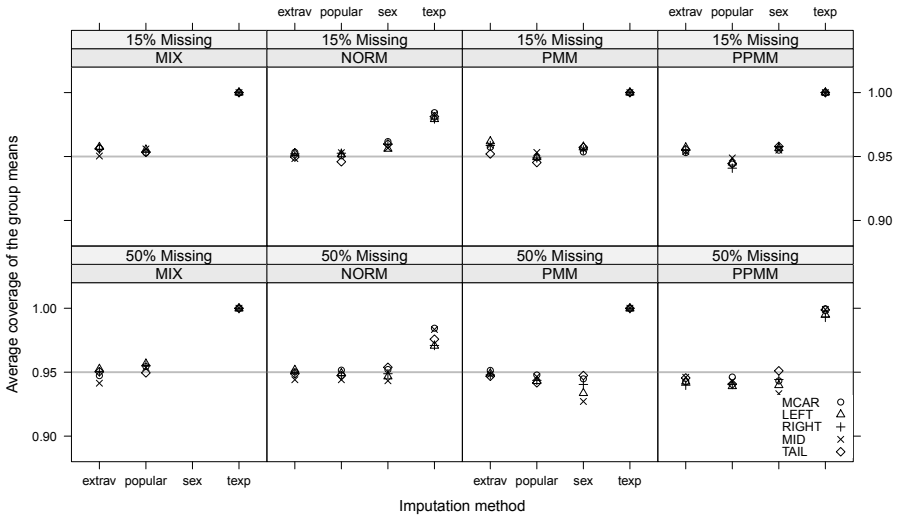

Figure 2:
Average coverage rate of the 95 percent confidence interval of the group means. Shown are results for four imputation approaches and four variables for varying missingness percentages.

Because teacher experience is a level-2 variable, values are the same for every pupil in a cluster. For such variables, PMM will yield correct results, even when none of the cluster's values are observed. MIX will also yield correct results, but only for clusters that have at least one observed value.

Coverage rates of the cluster means

Figure 2 displays the average over the coverage rates for the cluster means. It can be seen that performance for all methods is very stable across all variables, with very good coverage rates under all missingness mechanisms when considering 15 percent missingness. Because of the unbiased group means, PMM is able to perfectly cover the average teacher experience in the population after imputation, even for large missingness percentages.

When missingness increases to 50 percent, PMM performance is still good for all

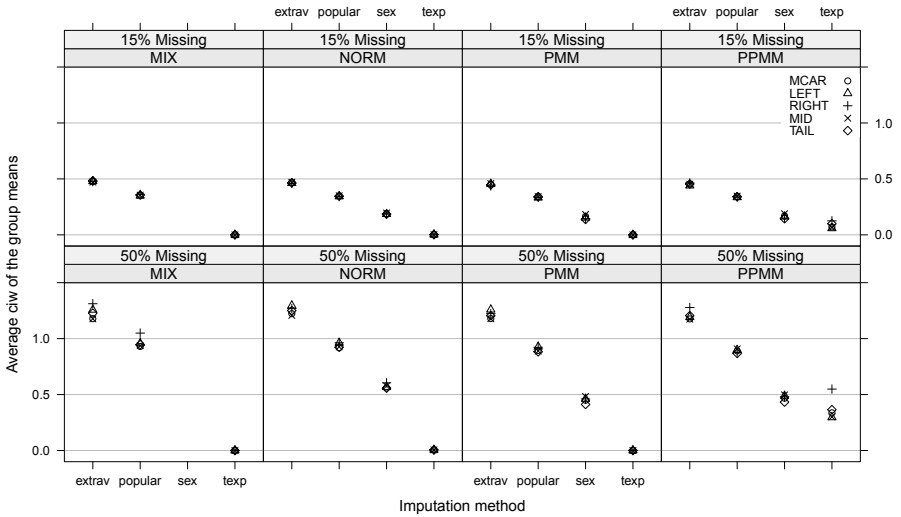

Figure 3:
Average width of the 95 percent confidence interval of the group means. Shown are results for four imputation approaches and four variables for varying missingness percentages.

variables. However, the dichotomous variable gender may be more efficiently imputed using logistic regression imputation (as under NORM).

Confidence interval width

Confidence interval widths are generally small when the cluster structure is taken into account, with PMM yielding slightly smaller intervals than the other imputation approaches. As expected, the average interval width for teacher experience under PMM is zero and the average interval width under NORM is close to zero. For MIX, the average interval width is unbiased, but is calculated over the observed clusters for large amounts of missingness. As expected, interval widths between PMM and PPMM are very similar, with the exception of the interval widths for experience of the teacher.

Table 3:
Variables in the SBD dataset

variable name	level	description	%mis
school_ID	2	school identifier	0 %
jurisdiction	2	school jurisdiction	0 %
sector	2	school sector	0 %
geolocation	2	school geographical location	2.21 %
sex	1	pupil gender	0.37 %
indigenous_status	1	pupil indigenous status	1.72 %
year_level	1	pupil year level	0.76 %
parent1_educ_schl	1	parent 1 school education level	17.30 %
parent1_educ_nonschl	1	parent 1 non-school education level	16.96 %
parent1_occ	1	parent 1 occupation	22.58 %
parent2_educ_schl	1	parent 2 school education level	29.92 %
parent2_educ_nonschl	1	parent 2 non-school education level	29.35 %
parent2_occ	1	parent 2 occupation	31.97 %

Application

We apply PPMM on a dataset collected by ACARA for the purpose of providing fair and meaningful comparisons of student performance in the National Assessment Program - Literacy and Numeracy (NAPLAN) between schools serving students from statistically similar socio-educational backgrounds. The resulting student background dataset (SBD) contains 2.782.060 Australian students clustered in 9.671 Australian schools and can be used for obtaining an estimate of the social educational advantage (SEA) score for Australian schools. An overview of the most important variables in the dataset is given in Table 3.

All variables with missingness are imputed, but here we focus on parent education and occupation. The parental variables are ordered categorical variables, with 'occupation' and 'non-school education level' having a separate category that records 'not in paid work' and 'no non-school qualification', respectively (see Table 4). The dual (or semi-categorical) nature of these data can be split in two parts: an ordered distribution over the categories and a point mass that does not follow the ordering of the other categories. For continuous or integer data with such a point mass, PMM is known to be a very effective single-step imputation approach (Vink et al., 2014).

Another reason for focusing on the parent variables is the large amounts of missingness. The parent variables contain most of the missing values in the data, ranging from 17 to

Table 4:
Levels of the parent variables in the SBD

Parent occupation
Senior management and qualified professionals
Business managers and associate professionals
Tradesmen/women, clerks and skilled staff
Labourers and related workers
Not in paid work in last 12 months
School education level
Year 12 or equivalent
Year 11 or equivalent
Year 10 or equivalent
Year 9 or equivalent or below
Non-school education level
Bachelor degree or above
Advanced diploma/Diploma
Certificate I to IV (including trade certificate)
No non-school qualification

32 percent missingness per variable. As a result, a large number of observations may be missing on the school level and schools are sometimes even completely unobserved. With such large amounts of missingness, it is important to use an imputation procedure that is able to capture the multilevel structure. Neglecting the clustering of the data will result in an underestimation of the intraclass correlation.

Finally, the parent variables are critical in the direct estimation of the SEA score for a school. See Acara (2014) for a detailed explanation of the modeling of Australian social educational advantage measures.

Procedure

We partitioned the SBD data into 271 parts based on jurisdiction, sector and geolocation, with parts containing only whole schools and schools not being split among parts. Each partition contains approximately 10,000 cases. Imputations are performed using PPMM with 5 imputations and 10 iterations for the algorithm to converge. We set $l = 1,000$

Table 5:
Intraclass correlations and average group means in the observed and imputed data. Shown are the average imputation value ($\hat{m}$) and the observed (but incomplete) data estimate (obs) for education (schooled and non-schooled) and occupation for both parents.

	ICC		MEANS	
	$\hat{m}$	obs	$\hat{m}$	obs
parent1_educ_schl	0.19	0.17	3.17	3.17
parent1_educ_nonschl	0.05	0.05	6.77	6.76
parent1_occ	0.18	0.17	4.36	4.32
parent2_educ_schl	0.20	0.18	3.02	3.04
parent2_educ_nonschl	0.05	0.06	6.56	6.55
parent2_occ	0.24	0.22	3.20	3.16

as the sample size for the random subset of observed donor candidates to speed up the imputation process.

Results

After imputation, the intraclass correlation in the parent variables is similar to the intraclass correlation of the parent variables in the incomplete data (see Table 5), with difference being generally very small. This indicates that the utilized imputation method was able to give meaningful predictions, thereby taking group membership into account.

Table 5 also shows the fixed effect estimates of the parent variables before and after imputation. It can be seen that difference is very small, when compared to the observed values, indicating that PPMM is able to very accurately estimate fixed effect from the incomplete data.

We used a random effects model to estimate social educational advantage in each of the imputed datasets. The model takes the form $SEA_{ab} = \mu + U_a + W_{ab}$ where SEA_{ab} is the score of the bth pupil at the ath school, μ is the overall average, U_a is the school-specific random effect and W_{ab} is the individual-specific error. The estimates and variances from the five imputed datasets were combined to obtain a Bayesian estimate for social educational advantage on the school level.

In Figure 4, we compare the conditional means from the random effects model to the group means from the fixed effects model. Note that the data has been sorted based

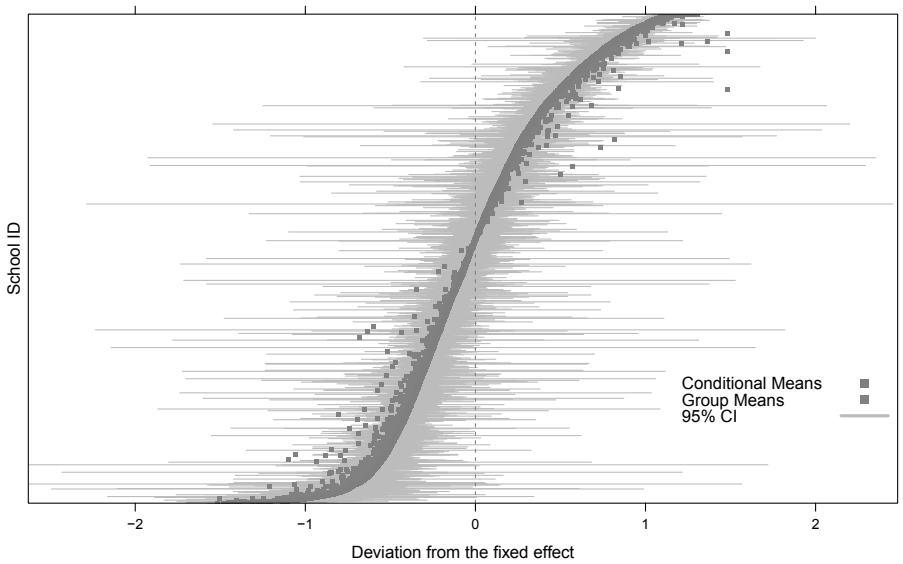

Figure 4:
Conditional SEA means from the random effects model and group SEA means from the fixed effects model compared after imputation. Shown are pooled results for the conditional means, the group means and the 95 % confidence interval for the conditional means.

on size of the conditional means. It can be clearly seen that the random effects model takes the group size into account and that shrinkage towards the fixed effect is applied. The larger confidence interval widths belong to the smaller schools that are completely unobserved. As a result of the increased uncertainty caused by the large amounts of missingness in these schools, the between imputation variance and, naturally, the confidence interval width increases.

Discussion

PMM emerges as a very effective imputation technique for multilevel data when the cluster structure is taken into account. The algorithm is able to preserve the multilevel

nature of the data, leading to precise and well-covered estimates.

Controlling for cluster membership is not the only requirement for a good multilevel imputation approach. In our view an imputation method for multilevel data must adhere to the following properties:

1. *Structure preserving:* The cluster structure should be accounted for during imputation.
2. *Generality:* The cluster size and the amount of missingness may vary and clusters may be completely unobserved.
3. *Observed plausibility:* Imputed values must be within the range of plausible values such that only realistic values can be imputed.

In a multilevel setting it can be concluded that PMM performance is comparable to dedicated methods for multilevel data and that PMM is sometimes even able to outperform dedicated methods, especially when the amount of missingness is large or when some clusters are completely unobserved.

For small cluster sizes and 50 percent missing data, there can be a slight (but conservative) overestimation of the ICC. However, in all simulation conditions PMM yields realistic imputations that are within the bounds of the plausible data values. In practice, this proves to be especially convenient when imputing continuous ratio scales, dichotomous variables, categorical variables, or even semicontinuous data.

We proposed partitioned predictive mean matching as a straightforward extension to the PMM algorithm that divides the big-data multilevel problem into manageable parts that can be solved by standard predictive mean matching. We have demonstrated that PPMM performance is similar to the performance of unpartitioned PMM, proving PPMM to be an effective imputation approach for large datasets, especially those datasets where dummy coding strategies are computationally not feasible.

The continuous variables in the simulation study are normally distributed. It is well known that deviations from normality can have a serious impact on the performance of methods that assume such distributions (MIX, NORM): evaluating the performance of these methods on non-normal data would be pointless. PMM, on the other hand, is known to handle deviations from normality very well (Vink et al., 2014). Our application on real data also demonstrates that PMM can still be used in situations where non-normal distributions are considered.

PMM is widely recognized as a method that preserves data distributions and, although it uses underlying methodology that assumes variables to be continuous, we have shown that it can yield plausible inference for continuous, ordered categorical, or even dichotomous multilevel data.

References

Acara (2014). *Guide to understanding 2013 index of community socio-educational advantage (ICSEA) values*. Sydney, Australia.: Australian curriculum assesment and reporting authority. Last accessed on Sep 30, 2014.

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal*, *53*(1), 57–74.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). *Handbook of statistical data editing and imputation*. Wiley.

Graham, J. W. (2012). *Missing data: Analysis and design*. Springer.

Heeringa, S., Little, R., and Raghunathan, T. (2002). *Survey Nonresponse*, chapter Multivariate Imputation of Coarsened Survey Data on Household Wealth, (pp. 357–371). Wiley.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2 ed.). Psychology Press.

Koller-Meinfelder, F. (2009). *Analysis of Incomplete Survey Data–Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching*. PhD thesis, Otto-Friedrich-Universität Bamberg.

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, *6*(3), 287–296.

Morris, T. P., White, I. R., and Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, *14*(1), 75.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, *4*(1), 87–94.

Schenker, N. and Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis*, *22*(4), 425–446.

Siddique, J. and Belin, T. (2007). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in medicine*, *27*(1), 83–102.

Su, Y., Gelman, A., Hill, J., and Yajima, M. (2011). Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statistical Software*, *45*(2).

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, *76*(12), 1049–1064.

Van Buuren, S. and Groothuis-Oudshoorn, C. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

Vink, G., Frank, L. E., Pannekoek, J., and Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, *68*(1), 61–90.

Vink, G. and Van Buuren, S. (2014). Pooling multiple imputations when the sample happens to be the population. arXiv:1409.8542 [math.ST].

White, I., Royston, P., and Wood, A. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399.

Yu, L., Burton, A., and Rivero-Arias, O. (2007). Evaluation of software for multiple imputation of semicontinuous data. *Statistical Methods in Medical Research*, *16*(243).

Zhao, J. H. and Schafer, J. L. (2013). *pan: Multiple imputation for multivariate panel or clustered data*. R package version 0.9.