Confidentialising Microdata Using Multiple Imputation:

Development and Evaluation of a Non-parametric Hierarchical Bayesian Imputation Model for Numerical Data

Patrick Graham
Senior Researcher, Bayesian Research,
Senior Research Fellow, Department of Public Health & General Practice,
University of Otago, Christchurch
patrick.graham@xtra.co.nz
patrick.graham@otago.ac.nz

Lena Rodnyanskiy
Statistical Analyst, Statistics New Zealand

Lisa Henley
Statistical Analyst, Statistics New Zealand
Manager, Analytical R&D, Groupe Aeroplan London

This report was commissioned by Official Statistics Research, through Statistics New Zealand. The opinions, findings, recommendations and conclusions expressed in this report are those of the author(s), do not necessarily represent Statistics New Zealand and should not be reported as those of Statistics New Zealand. The department takes no responsibility for any omissions or errors in the information contained here.
Abstract
When developing data products for use by external researchers, statistical agencies seek to meet legal and ethical commitments to respect respondent confidentiality by modifying the data originally collected. Statistical agencies also need to ensure that data released to researchers can be relied upon to produce valid inferences across a broad range of statistical analyses. To some extent these aims are in competition because modifying data before release may distort some inferences obtained from the released data. Multiply imputed synthetic data has been proposed as a solution to the data release problem faced by statistical agencies. In this report we summarise the background to the multiple imputation proposal and develop new multiple imputation methods for producing synthetic versions of data comprising a mix of categorical and numerical variables. The imputation models developed build on an earlier project which proposed hierarchical Bayesian imputation models for categorical data. This framework is extended to data comprising both categorical and numerical variables by modelling the joint distribution of numerical variables conditionally on categorical variables. Hierarchical Bayesian models for numerical variables are extended by using ideas and models from non-parametric Bayesian statistics. The resulting imputation method uses a generalised Polya urn sampling scheme which creates synthetic data as a mix of model-generated records and records sampled from the data. The performance of the proposed new imputation methodology is evaluated in two sets of comparisons. In the first, the new imputation method is compared with fully parametric hierarchical model based imputation in an exercise which uses the CURF for the 2003 Household Income Survey, as the real data to be confidentialised. In the second, the original 2003 Household Income Survey data is used as the real data and the performance of multiply imputed synthetic versions of this dataset, the CURF for this dataset and another confidentialising method, known as sufficiency based perturbation are compared.

Keywords
Confidentiality, hierarchical Bayes, multiple imputation, non-parametric Bayes, synthetic data.

Acknowledgements
The authors thank Mike Camden and Richard Penny for valuable technical advice and for sharing their knowledge of confidentiality issues in official statistics. Mike also acted as the project sponsor. Jim Young ported the PRIMM code from S+ to R.
Reproduction of material
Material in this report may be reproduced and published, provided that it does not purport to be published under government authority and that acknowledgement is made of this source.

Citation

Published by
Statistics New Zealand
Tatauranga Aotearoa
Wellington, New Zealand

ISSN 1177-5017 (online)
ISBN 978-0-478-31586-8 (online)
## Contents

1 Introduction  
1.1 A note on notation  

2 Multiply-imputed synthetic data: Theory.  
2.1 Modelling and imputation  
2.2 Inference given synthetic data  
2.3. Specification of imputation models for a mix of categorical and numerical variables: General considerations  

3. Imputation models for categorical variables: The role of hierarchical modelling  

4. Imputation models for the joint conditional distribution of numerical variables given categorical variables.  
4.1 Why constructing imputation models for numerical data is harder than for categorical data.  
4.2 A hierarchical multivariate normal imputation model.  
4.3. Limitations of hierarchical imputation models  
4.4. Non-parametric Bayesian models  
4.5. Combining hierarchical and non-parametric Bayesian models to model the conditional distribution of numerical variables given categorical variables.  

5. Fitting hierarchical and non-parametric Bayesian models  
5.1 Fitting hierarchical Bayesian Poisson log-linear models  
5.2 Fitting the hierarchical non-parametric Bayesian model for numerical variables  
5.2.1 Exploiting the Polya urn representation of the Dirichlet Process.  
5.2.2 A Gibbs sampling algorithm  

6. Generating multiply imputed synthetic datasets from the combined categorical and numerical data models.  

7.1. Data source and methods  
7.1.1 Data  
7.1.2 Imputation models.  
7.1.3 Analysis models.  
7.1.4 Model-fitting.  
7.1.5 Presentation of results.  
7.2 Descriptive statistics.  
7.3 Log-linear regression analysis models.  
7.3.1. Congenial models  
7.3.2 Models that are uncongenial due only to the inclusion of non-linear component of the age-sex interaction.  
7.3.3. Models that are uncongenial due only to differences in the scale on which models are specified.  
7.3.4. Models that are uncongenial due to differences in the scale used for model specification and inclusion of non-linear components of the age-sex interaction term.  
7.4 Disclosure risks.  

8.1 Design of the comparison.  
8.2 Descriptive statistics  
8.3. Log-linear regression models.  
8.3.1 Congenial Models.  
8.3.2. An analysis model that is uncongenial due solely to the inclusion of non-linear components of the age effect in age by sex interactions.  
8.3.3 Models that are uncongenial due only to differences in the scale on which models are specified.  
8.3.4 Models that are uncongenial due to differences in the scale used for model specification and inclusion of non-linear components of the age-sex interaction term.  

9. Discussion  

References  

Official Statistics Research Series, 5, 2009  
www.statisphere.govt.nz
Tables
Subgroup sample sizes and quartiles for weekly income in the CURF dataset. 28
Quartiles for hours worked per week for the overall CURF dataset and for 36 subgroups 29
Description of analysis models 31
Distribution of median prediction errors for 54 unique individuals under different synthetic data methods. 58
Proportion of synthetic data predictions of weekly income falling within specified bounds of actual values for 54 uniques. 58
Summary statistics for credible interval overlap based data utility for the CURF, synthetic and perturbed data under three uncongenial scenarios. 82

Figures
Figure 1 35
Figure 2 36
Figure 3 37
Figure 4 38
Figure 5 40
Figure 6 41
Figure 7 43
Figure 8 45
Figure 9 47
Figure 10 48
Figure 11 49
Figure 12 51
Figure 13 53
Figure 14 54
Figure 15 56
Figure 16 57
Figure 17 62
Figure 18 63
Figure 19 65
Figure 20 66
Figure 21 68
Figure 22 69
Figure 23 71
Figure 24 73
Figure 25 75
Figure 26 77
Figure 27 79
Figure 28 80
Figure 29 81
1 Introduction

Statistical agencies seeking to release data to external users face a tension between protecting respondent confidentiality and releasing data of sufficient detail and quality that analysts can trust the results obtained from analysis of the released data. Multiply imputed synthetic data was proposed by Rubin (1993) as one solution to this problem. In the multiple imputation framework an agency uses the originally collected data to develop a model from which multiple synthetic versions of the data are generated. In the case of fully synthetic data, no original data is released and the synthetic data records are draws from the posterior predictive distribution of new data conditional on the observed data and the imputation model. Disclosure risks for fully synthetic data are much reduced compared to the release of original data records since, being predictions from a model, the records in synthetic datasets cannot be linked to individuals in the original data. A variant of fully synthetic data generates synthetic data values only for certain variables, either those regarded as sensitive or variables that can be used to identify individuals (Abowd & Woodcock, 2001; Kennickel, 1997; Dreschler et al, 2008).

If conventional statistical models, such as generalised linear models, are used as imputation models, modelling assumptions such as absence of interaction or specifications for functional forms of relationships between variables, will be replicated, on average, in synthetic data. If the assumptions of the imputation model are erroneous and differ from those of the analysis model, synthetic data may fail to faithfully reproduce the estimated model the analyst would obtain if permitted to access the original data. It is therefore imperative to develop imputation models that produce synthetic data which exhibit a degree of robustness to mis-specification of the imputation model relative to the true data distribution and to the assumptions of the analyst’s model. In previous papers we have investigated the use of hierarchical Bayesian models as imputation models for categorical data and found that these models produce marked gains in robustness to prior imputation model mis-specification compared to conventional non-hierarchical models for categorical data such as conventional Poisson log-linear models (Graham, Young & Penny, 2008, 2009).

In this report we extend our hierarchical imputation modelling framework to datasets containing a mix of numerical and categorical variables by modelling the joint distribution of numerical variables conditionally on the categorical variables. We specify a hierarchical model for the regression of numerical variables on categorical variables and embed this hierarchical structure in a non-parametric Bayesian model (Muller & Quintana, 2004) in order to further improve robustness to mis-specification of the prior imputation model. Our imputation model results in a posterior predictive distribution which can be simulated by a generalised Polya urn and this leads to synthetic datasets which comprise a mix of original data records and model generated records. Although the resulting synthetic data could be viewed as partially synthetic because it contains some real data, its genesis is entirely within the theory of fully synthetic data and each synthetic record is a draw from the posterior predictive distribution for new data. Consequently, the inference framework for fully synthetic data (Reiter 2002; Raghunathan, Reiter & Rubin, 2003) applies rather than the corresponding framework for partially synthetic data (Reiter, 2003) which typically applies to situations involving a priori specification of a subset of variables or individuals to be imputed.

The structure of this report is as follows. In section 2 we briefly review the theory of multiply imputed synthetic data. In section 3 we review the use of hierarchical Bayesian models for categorical data and extend this framework to mixed numerical and categorical datasets in section 4 by combining hierarchical and non-parametric Bayesian models. In section 5 we discuss methods for fitting the proposed non-parametric hierarchical model and discuss generation of synthetic data from the non-parametric hierarchical imputation model in section 6. Sections 7 and 8 contain results of two empirical evaluations of the proposed imputation model. The results of these investigations are summarised and implications discussed in section 9 which also contains a discussion of possible avenues for extending and improving the imputation model.

1.1 A note on notation

We use the name of a distribution or standard abbreviation, such as MVN for multivariate normal, in upper case to refer to the distribution of a random quantity, whereas a distribution or standard abbreviation written in lower case refers specifically to a density (or probability mass function). Thus, for example, \( X \sim \text{POISSON}(\lambda) \) indicates \( X \) is distributed as a Poisson random deviate implying \( p(X) = \text{poisson}(X \mid \lambda) = \frac{1}{X!} \lambda^X \exp(-\lambda) \).
2 Multiply-imputed synthetic data: Theory

2.1 Modelling and imputation

Multiply imputed synthetic datasets are draws from the posterior predictive distribution of new data given the observed data. The new data can be thought of as the responses for a new sample of individuals, not included in the observed sample but drawn from the same population as the observed sample. Clearly these responses are unobserved and consequently multiple imputations are required to correctly represent the uncertainty concerning the response of the non-sampled individuals.

Imputations are created from a statistical model for the data, which is fitted to the observed data and then used to generate the new data. In standard parametric modelling set-ups the form of the model, in terms of both distributional assumptions (eg normal, binomial, Gamma, etc) and the functional form of regression relationships between variables (eg, linear, log-linear, logit etc) is assumed known, except for unknown parameters, such as regression parameters and measures of dispersion. Fitting the model to the data yields information about the unknown parameters but since the distributional assumptions and functional form of regression relationships are assumed known they are not modified by data and are assumed to hold also for predicting future data. Monte Carlo approximation of posterior predictive distribution then follows by repeating the following steps:

1. Draw the parameter vector from its posterior distribution
2. Draw new data from the assumed data-model with the parameter vector set to the value drawn in step (1).

These two steps constitute multiple imputation of a new sample, given the posterior distribution of the parameter vector of the assumed data model and define the generation of multiply imputed synthetic data under standard parametric modelling assumptions. Repeating this cycle $M > 1$ times generates $M$ multiply imputed datasets.

Mathematically, the posterior predictive distribution for new data, $D_{\text{new}}$, given an assumed data model, $F$ (henceforth the imputation model), the form of which is assumed known except for a vector of parameters, $\theta$, which are to be estimated from data, $D_{\text{obs}}$, can be written

\[
p(D_{\text{new}} \mid D_{\text{obs}}, F) = \int p(D_{\text{new}} \mid D_{\text{obs}}, F, \theta) p(\theta \mid D_{\text{obs}}, F) d\theta.
\]

In this notation, $F$ denotes the assumptions concerning both the distributional form of the data and the functional form of relationships between variables being modelled, such as the form of the regression relationship between outcome and predictor variables. If $F$ is assumed known it is not itself a random entity but is part of the prior information brought to the analysis and is therefore properly represented as part of the conditioning information set for the analysis.

In standard formulations of parametric inference and prediction, the parametric modelling assumptions are left implicit in expressions for posterior and predictive distributions. However we have explicitly included the modelling assumptions in the above formulation to make explicit the contrast between the standard parametric modelling set-up and the approach developed Section 4, in which the form of the imputation model, $F$, is regarded as uncertain and hence assigned a prior distribution.

The two-step Monte Carlo algorithm given above is a Monte Carlo approximation to the integral in (1), and, reading the integrand from right to left, the two Monte Carlo steps correspond to the two terms in the integrand.

In standard parametric model formulations, individual data-records are assumed conditionally independent given the model parameters. This implies the conditional distribution for $L$ records, denoted by $D$, generated from the model $F$, with parameter vector $\theta$, has the form:

\[
p(D \mid F, \theta) = \prod_{i=1}^{L} p(D_i \mid F, \theta)
\]

and, consequently, for observed data $D_{\text{obs}} = D_1, \ldots, D_n$, and new data $D_{\text{new}} = (D_{n+1}, \ldots, D_L)$, the posterior predictive distribution for $D_{\text{new}}$ given $D_{\text{obs}}$ is given by

\[
p(D_{\text{new}} \mid D_{\text{obs}}, F) = \prod_{j=n+1}^{L} p(D_j \mid F, \theta) p(\theta \mid D_{\text{obs}}, F) d\theta
\]
where:

\[ p(\theta \mid D^\text{obs}, F) \propto p(\theta \mid F) \prod_{i=1}^{n} p(D_i \mid \theta, F). \]

The number of imputations required in order for synthetic data to have good inferential properties appears to be considerably more than is required in conventional missing data problems in which the fraction of data which is missing is typically low (Graham, Young, & Penny 2008, chapter 1).

2.2 Inference given synthetic data

Users can obtain inferences from multiply imputed synthetic data by running analyses on each imputed dataset and then combining results across the imputations using formulae derived by Raghunathan, Reiter, Rubin (2003). The combining formulae assume a t-approximation to the posterior distribution for unknowns given the synthetic data. These formulae differ from those used for combining inference across multiple imputations in standard missing data problems (Rubin, 1987) and also differ from the combining formulae for partially synthetic data (Reiter, 2003).

2.3 Specification of imputation models for a mix of categorical and numerical variables: General considerations

Imputation models are models for the joint distribution of all variables for which imputations are required. In Graham, Young & Penny (2008) we considered the case where all variables of interest are categorical, implying that the joint distribution can be represented by a multiway table of counts cross-classifying the categorical variables. Modelling the cell-counts of the multiway table is therefore equivalent to modelling the joint distribution of the categorical variables.

We used hierarchical Bayesian Poisson log-linear models to model the cell counts and found that the synthetic data generated using this modelling framework had good inferential properties compared to synthetic data generated from non-hierarchical models (Graham, Young & Penny, 2008, 2009). Many datasets contain a mix of categorical and numerical variables and our approach for modelling categorical data therefore needs to be extended to accommodate more general datasets. In view of the success of the hierarchical Bayes Poisson log-linear model framework for categorical data, it is natural to extend the imputation model to numerical variables by modelling the numerical variables conditionally on the categorical variables.

More formally, we consider that variables can be one of two types, categorical, denoted by \(X\), or numerical, denoted by \(Y\), and propose to model the joint distribution of all variables by:

\[ p(Y, X \mid \theta) = p(Y \mid X, \theta_y) p(X \mid \theta_x) \quad (2) \]

where \(\theta = (\theta_y, \theta_x)\) and both \(Y\) and \(X\) are multivariate. If the parameters of the models for the categorical variables and the numerical variables given the categorical variables are a priori independent then it follows that the joint posterior distribution of the model parameter given observed data \((Y^\text{obs}, X^\text{obs})\) is given by:

\[ p(\theta \mid Y^\text{obs}, X^\text{obs}) = p(\theta_y \mid Y^\text{obs}, X^\text{obs}) p(\theta_x \mid X^\text{obs}). \]

The practical implication of this a posteriori independence is that, under standard parametric imputation models, imputations of new data are drawn by:

Drawing \(\theta_y\) and \(\theta_x\) independently from their respective posterior distributions.

Generating a new set of categorical data by drawing data from the categorical data model \(p(X \mid \theta_x)\), with the parameter \(\theta_x\) set at the value drawn in step (I).

Drawing numerical data, conditionally on the categorical data generated in step (II) by drawing from \(p(Y \mid X, \theta_y)\), with the parameter \(\theta_y\) set at the value drawn in step (I).

Our use of the terms ‘numerical variables’ and numerical data is meant to include all ordered variables which cannot be sensibly modelled as categorical. This includes variables which are not strictly continuous, because of rounding, but which nevertheless are not convenient to treat as categorical because of the number of possible values which they can take. Recorded income is a
case in point because this variable is often rounded to a dollar value (or nearest $10) but is usually modelled by continuous data models such as log-normal or gamma distributions.
Turning now to the construction of imputation models for multivariate categorical data, we first note the corresponding non-hierarchical imputation models. That in virtually all analytical situations multivariate categorical data can be completely described by a multi-way table of cross-classified cell counts. Given the cell counts, the individual unit records of categorical variables generally contain no additional information about the joint distribution of the categorical variables involved in the cross-classification. For example, in standard sampling situations order of observation within a cell is rarely informative and observations within a cell are usually judged exchangeably. If, for some reason, order of observation within a cell is judged to be informative with respect to the joint distribution of categorical variables then a multi-way table may not be the best representation of the data unless observation order can be grouped rather coarsely and added to the cross-classification.

Circumstances where individual categorical data records within a cell contain information about the joint distribution of the cross-classified categorical variables, additional to that contained in the cell-count, seem very rare and would usually indicate that an additional variable is needed to properly describe the data. Consequendy in constructing an imputation model for categorical data it will virtually always be sufficient to model the cell-counts of the multi-way table cross-classifying the categorical variables. Predictions from a model for cell counts of a multi-way table will produce sets of synthetic cell counts which can be converted to the corresponding numbers of unit-records if required.

Poisson log-linear models are a popular and convenient approach to modelling cell counts in a multi-way table. The use of hierarchical Bayesian Poisson log-linear imputation models for generating synthetic categorical data is investigated in Graham & Penny (2007) and Graham, Young & Penny (2008). In the latter paper the advantage of hierarchical imputation models for synthetic data compared to conventional non-hierarchical models is illustrated in simulation studies which show that synthetic data generated under hierarchical imputation models are considerably more robust to discrepancies between the imputer’s prior model structure and the analyst’s model than synthetic data generated under conventional non-hierarchical log-linear imputation models. Here we recap only the main ideas of hierarchical Bayesian Poisson log-linear models for categorical data, as they pertain to synthetic data. These ideas are a useful stepping stone to constructing imputation models for numerical data conditionally on categorical data, as discussed in section 4 below.

Let $C = \{C_j, j = 1, \ldots, J\}$ denote the J cell counts of the cross-classification of the categorical variables. Each cell is defined by a unique combination of values of the categorical variables (eg Age = 25 to 29, sex = male, ethnicity = Māori, educational level = tertiary, and so on) and we let $X_j$ be a
vector of indicators describing the cell. A hierarchical Bayes Poisson log-linear model for the cell counts can be specified as follows:

\[
C_j \mid X_j, \lambda_j \sim \text{POISSON}(\lambda_j), \quad \text{independently, for } j = 1, \ldots, J
\]

\[
\lambda_j \mid X_j, \beta, \xi \sim \text{GAMMA}(\xi, \xi / \eta_j),
\]

\[
\ln(\eta_j) = H_{\text{cat}}(X_j)\beta
\]

where \( H_{\text{cat}}(X_j) \) is the expansion of the cell–descriptor vector \( X_j \) into a \( 1 \times p \) ‘design vector’ which includes an intercept term and may include transformations of the elements of \( X_j \) and product terms involving combinations of the individual elements of \( X_j \), or transformations of these elements. For example, considering a simple example involving only age category and sex the cell corresponding to males aged 25 to 29 could be indicated by \( X_j = (25,1) \) whereas for a model with an intercept term, age coded linearly and an age by sex interaction term, the design vector representing this cell would be \( H_{\text{cat}}(X_j) = (1, 25, 1, 25) \). Further terms could be added if non-linear age effects are to be modelled. The ‘cat’ subscript is used to emphasise that this design vector is specific to the categorical data model. Modelling of the conditional distribution of numerical variables given the categorical variables may require a different design vector representation of a cell in order to reflect the modelling assumptions for regressing the numerical variables against the categorical variables.

The model specification is completed by specification of priors for the hyper-parameters, \( \beta \) and \( \xi \). Under the gamma model of (3), the prior mean of the cell-specific Poisson parameters is

\[
E(\lambda_j \mid X_j, \beta, \xi) = \eta_j = \exp(H_{\text{cat}}(X_j)\beta)
\]

and the prior variance is

\[
V(\lambda_j \mid X_j, \beta, \xi) = \eta_j^2 / \xi.
\]

The distinction between the hierarchical and conventional, non-hierarchical Poisson log-linear models is equation (3). In the hierarchical model specification the Poisson parameters are modeled as draws from cell-specific gamma distributions whereas in a conventional log-linear model the Poisson parameters would be modeled as deterministic functions of the cell descriptors, \( X_j \). In the hierarchical model this deterministic model is used, not for the Poisson parameters, but for the prior mean of these parameters and is therefore a statement of prior expectation of the relationship of the cell-specific Poisson parameters to the cell descriptors. The hierarchical model allows for uncertainty about the form of the model linking the Poisson parameters to the cell descriptors, by embedding the model for this relationship in a probability model; the gamma model of equation (3). In contrast specifying a conventional non-hierarchical log-linear models is equivalent to asserting that the log-linear model linking Poisson parameters to cell descriptors holds with probability one.

Under the prior Gamma model of (3), the conditional (on \( \beta, \xi \)) posterior distribution for the cell means is

\[
\lambda_j \mid C^{\text{obs}}, \beta, \xi \sim \text{GAMMA}(\xi + C_j, \xi / \eta_j + 1)
\]

which has expectation

\[
E(\lambda_j \mid C, \beta, \xi) = B_j \eta_j + (1 - B_j)C_j,
\]

where \( B_j = \xi / (\xi + \eta_j) \)

(Gelman et al, 2004, pp. 51-54).
Equation (6) has the form of a weighted average of the observed cell count, $C_j$, and the prior mean $\eta_j = \exp(H_{cat}(X_j)\beta)$. Thus, the posterior mean shrinks the observed cell count towards the prior mean by an amount determined by the shrinkage parameter, $B_j$, which is a function of the hyper-parameters $(\beta, \xi)$. As the precision hyper-parameter, $\xi$, increases, so does the amount of shrinkage towards the prior mean. In the limit as $\xi$ tends to infinity the Poisson parameters are shrunk completely to the log-linear model specification and the hierarchical model collapses to the conventional non-hierarchical Poisson log-linear model.

The importance of (6) for synthetic data is that under a hierarchical Poisson (ie Poisson-Gamma) log-linear imputation model synthetic data will, on average, replicate the posterior expected cell counts which depend on both the observed cell counts and the prior model. Under a conventional non-hierarchical imputation model synthetic data will, on average, replicate the expected cell counts implied by the prior model. In practice, the hyper-parameters $(\beta, \xi)$ are not known and must be assigned prior distributions so that posterior distributions for these parameters can be obtained. Better fitting log-linear structures lead to the posterior for $\xi$ being located at larger values than is the case for poor fitting log-linear model structures. Hence a better fitting prior log-linear model will lead to more shrinkage towards the prior mean than a poor fitting prior model.

Assuming conditional independence of cell counts given the Poisson parameters, $M$ synthetic sets of cell counts can be generated by repeatedly:

Drawing from the posterior of the model hyper-parameters $p(\beta, \xi | C_{obs})$;

For $j = 1, \ldots, J$, draw cell-specific Poisson parameters independently from the GAMMA distributions given in (5);

For $j = 1, \ldots, J$, draw cell counts from independent Poisson distributions with Poisson parameters set to the values generated in step 2.
4 Imputation models for the joint conditional distribution of numerical variables given categorical variables

4.1 Why constructing imputation models for numerical data is harder than for categorical data

Constructing imputation models for the conditional distribution of numerical variables given categorical variables is more challenging than constructing imputation models for the joint distribution of categorical variables. As noted in section 3, in the categorical data case, except in very unusual circumstances, the cell counts of the multi-way cross-classification of variables are sufficient statistics for all analyses. The cell counts contain all the information in the data concerning the joint distribution of categorical variables and, given, the cell counts, the individual data records contain no additional information about this distribution. Consequently, any imputation model which yields predicted cell counts which are reasonable approximations to all observed cell counts will generate synthetic data which is a reasonable proxy for the real data, for all analyses. In practice it is difficult to obtain close approximations to all cell counts without fitting near-saturated models which are likely to be associated with unacceptable disclosure risks. Nevertheless from the shrinkage formula (6) it is clear that hierarchical Bayes imputation models will always yield synthetic data which on average approximates the observed cell counts more closely than synthetic data obtained from corresponding non-hierarchical model.

In the case of numerical data, the sufficient statistic analogous to the multiway table of cell counts for categorical data is the multivariate empirical distribution function. However, whereas Poisson log-linear and other categorical data models directly model multi-way cell counts, standard models for numerical data model only a limited number of features of the empirical distribution function such as the mean and variance, in the case of models based on the normal distribution. Embedding such models in a hierarchical structure can yield more accurate estimates of the mean of the underlying distribution but will not necessarily lead to improved estimation of higher order moments or finer features of the distribution. This point is explored further in 4.2 where we investigate the consequences of embedding the most common multivariate model for numerical data, the multivariate normal in a hierarchical model.

4.2 A hierarchical multivariate normal imputation model

Suppose a hierarchical multivariate normal regression model is used to model the joint conditional distribution of \( d \) numerical variables given categorical variables. Assuming, \( j = 1 \ldots J \), indexes cells in the multi-way cross-classification of categorical variables and \( i = 1, \ldots, C_j \) indexes individuals in the \( j \)th cell, this model can be written as follows

\[
[Y_j \mid X_j, \mu_j, V_j] \sim \text{MVN}(\mu_j, V_j), \quad i = 1, \ldots, C_j
\]

\[
\mu_j \sim \text{MVN}(Z_j \gamma, \Sigma), \quad j = 1, \ldots, J
\]

where, \( Z_j \) is a \( d \times dq \) matrix with the \( 1 \times q \) design vector \( H_{\text{num}}(X_j) \), defined analogously to \( H_{\text{cat}}(X_j) \) in the categorical data case, occupying elements \((k-1)q + 1\) to \(kq\) in row \( k \), \( k = 1 \ldots d \) and remaining elements of the matrix set to zero. For \( d = 2 \), \( Z_j \) therefore has the following form

\[
Z_j = \begin{bmatrix}
H_{\text{num}}(X_j) & 0_q \\
0_q & H_{\text{num}}(X_j)
\end{bmatrix}
\]

with \( 0_q \) denoting a \( 1 \times q \) zero vector. We are assuming here that the design vector and hence the regression relationship with the categorical variables is the same for each component of \( Y \). Extension to separate regression models is straightforward but introduces additional notation, so, as a simplification, we work with the assumption of a common regression model for all numerical variables.
The hierarchical multivariate normal model is completed by specification of a prior distribution for the hyper-parameters \((\gamma, \Sigma)\), which we discuss later.

As another simplification, we assume the cell-specific variance matrices \(V_j\) are known, as in Morris (1983), Everson and Morris (2000) and Graham (2005), though in reality they will be estimated from the data. Fixing the within cell variances simplifies model fitting (Everson and Morris, 2000; Graham, 2005). For creating synthetic data it may, in fact, be advantageous to fix the cell-specific first-stage variances at their observed values as this effectively conditions the analysis on the observed within cell variance and correlation structure. Synthetic data generated under a hierarchical model with first stage variances fixed at their observed values can be expected to more accurately reproduce within cell correlation structures than synthetic data generated under a hierarchical model with modeled first-stage variances.

Note that the hierarchical model differs from a non-hierarchical multivariate normal regression model by the adoption of the probability model (8) for the cell-specific means. This model reflects uncertainty in the linear model for the cell-specific means, with larger second-stage variances indicating more uncertainty. As the components of \(\Sigma\) shrink to zero the hierarchical model reduces to a non-hierarchical model.

Under the above hierarchical model specification the posterior expectation for the cell-specific means is given by

\[
E(\mu_j \mid Y, \gamma, \Sigma) = W_j \times Z_j \gamma + (I - W_j) \times \bar{Y}_j
\]

(9)

where

\[
W_j = \tilde{V}_j (\tilde{V}_j + \Sigma)^{-1}, \quad \bar{Y}_j = V_j / C_j, \quad \bar{Y}_j = (1 / C_j) \sum_{i=1}^{C} Y_{ij}.\]

(Lindley & Smith, 1972).

Thus, under the hierarchical model the posterior expectations for the cell means are compromises between the sample average and prior mean specified by the linear model \((Z_j \gamma)\). As the second stage (or prior) variance, \(\Sigma\) tends to zero the weight on the prior mean tends to one and the posterior expectation approaches the linear model estimate. This is consistent with the interpretation of \(\Sigma\) as a measure of uncertainty about the prior linear model. When there is no uncertainty \(\Sigma = 0\) and the hierarchical model collapses to a non-hierarchical model.

From the viewpoint of synthetic data the relevance of (9) is that, averaging over multiple imputations, cell means computed from synthetic data will approximate (9), rather than the model based estimate, \((Z_j \gamma)\), as would be the case if a non-hierarchical model was used as the imputation model. Since (9) is a weighted average of the observed cell means and the model-based estimate, synthetic data derived from a hierarchical imputation model can be expected to provide better estimates of cell means than non-hierarchical models. Consequently, all analyses for which the set of cell means can be regarded as sufficient statistics can be expected to be more nearly preserved under hierarchical imputation models than under non-hierarchical imputation models. However, even exact preservation of cell means would not ensure preservation of other features of the distribution of numerical variables, such as high-order moments, particular quantiles and so on. Consequently, while a hierarchical normal imputation model will improve on a non-hierarchical normal imputation in terms of yielding good synthetic data estimates of means and analyses which depend on means, performance for other characteristics of the distribution cannot be guaranteed and will depend on how accurately the distributional assumptions of the model reflect the data distribution.

4.3 Limitations of hierarchical imputation models

Standard distributional assumptions such as the normal, gamma or log-normal only approximate real data distributions and the choice of such models for particular applications often owes as much to convenience as to a thorough understanding of the distributional form from which the observed data is assumed to have been sampled. In standard generalised linear models it is only the mean response which is explicitly modelled as a function of covariates and hierarchical generalised linear models extend standard generalised linear models only by embedding the model for the mean response in a probability models. While this allows for uncertainty in the form of the mean-response model, the hierarchical model formulation still imposes specific distributional assumptions on the data. For
example, extending a Poisson model by adopting a gamma model for the Poisson parameters as in (3) implies a negative-binomial model for the cell counts. Adopting a normal model for the cell-means in normal regression model as in (9) implies a normal distribution for the individual responses but with an extra component of variance, compared to the corresponding non-hierarchical model.

The hierarchical normal model could be made more flexible in various ways. For example, the multivariate normal assumption is quite restrictive, implying linear relationships between the component variables as well as normality of the marginal distributions. These assumptions could be relaxed by noting that a set of \( p \) numerical variables \( Y = (Y_1 \ldots Y_p) \) can be modelled sequentially using

\[
p(Y | X, \theta) = p(Y_1 | X, \theta_1) p(Y_2 | Y_1, X, \theta_2) \ldots p(Y_p | Y_{p-1}, X, \theta_p).
\]

and choosing appropriate models for each component of the sequence, allowing non-linearity and non-normality as needed. If the implied distributional assumptions of the sequence of conditional models closely match the observed data distribution then such modelling will produce synthetic data which does a better job of preserving the real data than synthetic data created under the hierarchical normal model. However, in order to obtain models which can accommodate complex distributional forms additional flexibility beyond that provided by hierarchical generalised linear models may be required. One promising approach for constructing more flexible models is outlined in the following subsection and combined with hierarchical models in 4.5.

### 4.4 Non-parametric Bayesian models

Standard statistical models specify specific distributional forms which, typically, depend on a low dimensional vector of parameters. Under such models (eg the normal model with parameters mean and variance), inference for model parameters is equivalent to inference for the distribution because, given the model parameters the distribution is completely determined. An alternative modelling approach is provided by so-called non-parametric Bayesian models in which distributions are themselves treated as unknowns, and assigned a prior over the space of distributions (Walker et al, 1999; Muller & Quintana, 2004). The observed data updates the prior to a posterior distribution over the space of distributions. Conceptually the model can be represented as follows:

\[
Y_1, \ldots Y_n | F \sim F \\
F \sim p(F).
\]

where \( Y_1, \ldots Y_n \) denotes the \( n \) data vectors for a sample of size \( n \). From a synthetic data perspective, the posterior predictive distribution of new data is of prime importance rather than the posterior for the random distribution \( F \). In a non-parametric Bayesian framework, given observed data \( Y^{obs} \), the creation of \( M \) multiply imputed synthetic datasets can be represented by the following algorithm:

For \( m = 1 \ldots M \)

1. draw \( F_m \) from \( F | Y^{obs} \)
2. draw \( Y_{new}^{m} \) from \( F_m \).

For creating synthetic data the non-parametric approach has conceptual appeal. We use the observed data to learn about the distribution from which the observed data was sampled and then use this knowledge, encapsulated in \( F | Y^{obs} \), to generate new data by first sampling a distribution from \( F | Y^{obs} \) and then generating data from the sampled distribution. Fienberg, Makov & Steele (1998) proposed a similar conceptual framework for synthetic data but concentrated on categorical data.

Formally, under a non-parametric model the posterior predictive distribution for new data can be written

\[
p(Y_{new} | Y^{obs}) = \int p(Y_{new} | Y^{obs}, F) p(F | Y^{obs}) dF
\]

and under i.i.d. sampling of data from \( F \), this reduces to

\[
p(Y_{new} | Y^{obs}) = \int p(Y_{new} | F) p(F | Y^{obs}) dF
\]
which may be compared to the equivalent expression in the parametric case, given in equation (1). The integration over the posterior for \( F \) can be conceptualised in the Monte Carlo sense of drawing \( F \) from its posterior and this leads to the two-step Monte Carlo procedure given above. Nevertheless, drawing directly from a posterior over the space of distributions is a conceptually and practically challenging task. In the approach developed below explicit simulation of the posterior for the unknown distribution is avoided by an implicit integration over the posterior and this permits simulation from the posterior predictive distribution of new data without requiring explicit simulation of the posterior for \( F \).

A common theme in non-parametric Bayesian modelling is that the prior for the random distribution can be centered at a parametric distribution. This simplifies construction of the prior. One of the most popular non-parametric Bayesian models is the Dirichlet Process, which explicitly extends standard parametric forms by embedding them in a probability model as follows: The distribution \( F \) of the random variable \( Y \) (possibly vector-valued) is said to have a Dirichlet Process prior, centered on the parametric distribution \( G_\theta() \), with precision parameter \( \alpha \), if for any partition \( B = \{ B_1, \ldots, B_k \} \) of the sample space of \( Y \), the probabilities assigned by \( F \) to the elements of \( B \) have a Dirichlet distribution with parameters \( \{ \alpha G_\theta(B_1), \alpha G_\theta(B_2), \ldots, \alpha G_\theta(B_k) \} \). We use the notation \( F \sim DP(\alpha, G_\theta) \) to indicate that the random distribution is \( F \) is being modelled with a Dirichlet Process centered on \( G_\theta \).

For now we assume the parameters of the parametric distribution and the precision parameter, \( \alpha \), are known. We also restrict attention to modelling data for a single population or within a single cell of a multi-way stratification. These restrictions simplify the presentation and are relaxed in 4.5 below. The parameter \( \alpha \) is restricted to the positive real line and determines the weight given to the centering distribution \( (G_\theta) \) in the posterior for \( F \). The Dirichlet Process model restricts \( F \) to be discrete and the parameter \( \alpha \) also determines the 'clumpiness' or degree of clustering in the distribution (Escobar, 1994; Ohlssen et al, 2007). This latter characterisation of \( \alpha \) has been used by some authors (Escobar 1994, Ohlssen, et al, 2007) to define a prior for \( \alpha \), via a prior for the expected number of clusters. The centering distribution, \( G_\theta \), can be viewed as a prior guess at the form of \( F \) and from properties of the Dirichlet distribution it is immediate that for any element \( B_k \) of the partition, \( B \), \( E(F(B_k)) = G_\theta(B_k) \).

Although the restriction of the Dirichlet Process to discrete distributions causes no problem in the applications presented in sections seven and eight, it is generally regarded as a limitation of the model (Walker et al, 1999, Muller & Quintana, 2004). Another limitation is that the Dirichlet Process model restricts \( F \) to be discrete and the parameter \( \alpha \) also determines negative correlation between these probabilities. On the other hand, two practical advantages of the Dirichlet Process model are ease of posterior updating and the convenient form of the posterior predictive distribution for new data. Given data \( D^{obs} = \{ Y_1, \ldots, Y_n \} \), the posterior for \( F \) is defined by a Dirichlet Process with precision parameter updated to \( \alpha + n \), and centering distribution updated to a weighted average of the prior centering distribution, \( G_\theta \), and the empirical distribution of the observed data, \( F_n \). That is the posterior for \( F \) can be represented as follows

\[
F \mid D^{obs} \sim DP(\alpha + n, G_\theta^{post}),
\]

where \( G_\theta^{post} \) is defined as the mixture \( (\alpha G_\theta + nF_n) / (\alpha + n) \).

Another convenient feature of the Dirichlet Process model is that under this model the implied model for the data is a generalised Polya urn centered on the parametric form \( G_\theta \) (Blackwell & MacQueen, 1973). That is, the marginal distribution of the data, integrating over the random distribution \( F \), can be simulated using a generalised Polya-urn sampling model which proceeds as follows: The first observation is generated from the parametric model, \( G_\theta \) and a copy of this observation initialises a 'data-urn.' Then for \( i = 2 \ldots n \), the \( i^{th} \) observation is generated from the parametric model with probability \( \alpha / (\alpha + i - 1) \) and with probability \( (i - 1) / (\alpha + i - 1) \) is drawn randomly, with replacement, from the data-urn. A copy of the \( i^{th} \) observation, whether generated from the model or drawn from the data-urn, is added to the data-urn. If an observation is drawn from the data-urn, values in the data-urn are sampled with equal-probability. The generalised Polya urn (henceforth just Polya urn) model for the data is useful for deriving a marginal likelihood for the parameters of \( G_\theta \).
integrating over the random distribution, \( F \). This marginal likelihood is used in model-fitting as discussed in section 5, below. Moreover, and of particular relevance to synthetic data, the posterior updating of Dirichlet Process model to another Dirichlet Process, implies that given \( n \) observations, the posterior predictive distribution of a further \( m \) observations can also be described by Polya urn model with data-urn initialised with the \( n \) observed data points (Blackwell and MacQueen, 1973). That is, under the Dirichlet Process model the posterior predictive distribution for \( K \) new data records, given \( n \) observed records can be generated as follows: For \( i = 1 \ldots K \), draw from the parametric model with probability \( \alpha / (\alpha + n + i - 1) \) and, with probability \( (n + i - 1) / (\alpha + n + i - 1) \), draw, with replacement and equal probability, from the data urn. Each generated observation, whether drawn from the model or the data urn, is added to the data urn. (Walker et al, 1999).

In the context of creating synthetic data a major implication of the Polya urn sampling process is that it leads to synthetic datasets which are a mix of records drawn from the real data and records drawn from a parametric model. The ratio of real to model generated records depends on the relative size of the Dirichlet Process precision parameter, \( \alpha \) and the sample size \( n \). Large values of \( \alpha \), have the effect of increasing the frequency of model-generated records in the imputed synthetic datasets and large sample sizes have the effect of increasing the frequency of real data records in the synthetic datasets. However, it is important to note that as the Polya urn sampling process unfolds the data urn contains a mix of records drawn from the real data and model-generated records and therefore a draw from the data urn may contain a real data record or a model-generated record. This means that the number of real data records appearing in the imputed datasets is difficult to predict. The presence of some real data records in the imputed data sets can be expected to improve the concordance of inferences obtained from synthetic data and from real data. In principle, the presence of some real data records in synthetic data increases disclosure risk compared to fully model-generated synthetic data. However, given that the proportion of records which will be drawn from the data is unknown in advance, the random selection of records and the fact that a user would have no way of distinguishing real from model-generated records, the actual increase in disclosure risk may be negligible. Some comparative results on predictive disclosure risk are presented in section seven.

The preceding discussion of the Dirichlet Process and Polya urn sampling of predictive distributions discussion has treated the parameters \( \theta \) of the centering distribution \( G_{\alpha} \), as known whereas, in practice these will be unknowns which must also be assigned a prior distribution. Moreover, in order to keep things simple the presentation implicitly assumed a single group or cell. A more complete specification of the Dirichlet Process model in the context of modelling the conditional distribution of numerical variables given categorical variables, including specification of the centering distribution and priors for the parameters of that distribution is taken up in the following subsection.

### 4.5 Combining hierarchical and non-parametric Bayesian models to model the conditional distribution of numerical variables given categorical variables

Returning to the specific problem of modelling the conditional distribution of numerical variables given categorical variables, we note that cell-specific hierarchical models for numerical variables can be combined with non-parametric Bayesian models in two ways. Firstly, the usual assumption of a parametric distributional form for the conditional distribution of the numerical data, given the categorical variables, such as the multivariate normal assumption of (7), can be replaced by a non-parametric Bayesian model, as described in section 4.4 above. For example, the multivariate normal model in (7) could be replaced by a Dirichlet Process centered on the multivariate normal model. Alternatively, the parametric second-stage model for the cell means, such as the multivariate normal model in (8) could be replaced by a non-parametric model. Combining (7) with a non-parametric alternative to (8), would lead to a more flexible overall model for the numerical data, than the fully parametric model defined by (7) and (8) because this amounts to moving from multivariate normal mixing model for the cell means to a more flexible mixing model. However, under a hierarchical set up with a parametric first stage model (eg (7)) and non-parametric second stage model, the generation of synthetic data would not follow the Polya urn process outlined in the previous section because the Dirichlet Process is adopted for the cell-means rather than the data. Instead each synthetic dataset would be generated from the parametric first stage model, given cell-means, themselves generated from a posterior distribution based on the non-parametric model and the observed data.
In this report we concentrate on hierarchical models in which the usual parametric first stage distribution is replaced by a Dirichlet Process model. This means we can take advantage of the Polya urn model for the data, which simplifies model fitting (see section 5) and leads to the imputation process outlined in 4.4 which produces synthetic datasets comprising a mix of real and model-generated records. Fitting the model implied by the alternative approach of replacing the second stage model for cell-means by a Dirichlet Process, involves using a finite mixture approximation to approximate the Dirichlet Process model (eg Sethuraman, 1994; Ishwaran & James, 2001; Ohlssen et al, 2007). Our preliminary investigation of this method suggested that computation time for fitting such a model would be substantial and impractical for our standard, single processor computing environment.

In the remainder of this report we restrict attention to Dirichlet Process models for numerical data which are centered on a multivariate normal distribution. While other centering distributions could be considered the adoption of a multivariate normal centering distribution simplifies model-fitting by requiring only minor modification of a previously published Gibbs Sampling algorithm for fitting hierarchical multivariate normal models (Graham, 2005). A second reason for adopting a multivariate normal centering distribution instead of, for example, a sequence of conditional univariate distributions is that the former could be applied routinely by an agency considering release of synthetic data and wrapping the multivariate normal in Dirichlet Process model then provides a way of robustifying synthetic data against discrepancies between the actual data distribution and multivariate normality. The success or otherwise of this strategy is evaluated in sections 7 and 8 where we compare analyses based on synthetic data with analyses based on the underlying real data.

With the restriction to multivariate normal centering distributions, the non-parametric hierarchical model we consider for modelling the conditional distribution of numerical variables, given categorical variables can be represented as follows:

\[
Y_{ij} \mid X_j, V_j \sim F_j, \quad j = 1 \ldots J, \quad i = 1 \ldots C_j
\]

\[
F_j \sim DP(\alpha, G_{\theta_j}),
\]

\[
\theta_j = (\mu_j, V_j),
\]

\[
G_\theta \Leftrightarrow MVN(\mu_j, V_j)
\]

\[
\mu_j \mid X_j, \Sigma \sim MVN(Z_j, \gamma, \Sigma), \quad j = 1, \ldots, J
\]

For the applications discussed in section 7 and 8 we adopted a uniform prior for the regression hyper-parameters, \(\gamma\), which is a special case of a \(MVN(\gamma_{prior}, \Omega)\) as the diagonal components of \(\Omega^{-1} \rightarrow 0\). A finite variance, non-zero mean prior for the regression hyper-parameter, \(\gamma\), poses no additional computational difficulty. The Gibbs Sampler outlined in the following sub-section handles this more general prior, which provides a means of incorporating specific prior information on the direction or magnitude of regression relationships into the analysis. We adopt a multivariate uniform shrinkage prior for the second-stage variance matrix, \(\Sigma\) (Daniels, 1999). This prior is motivated by, and draws its name from, the idea of assigning a uniform prior to the weight given to the prior mean in shrinkage formulae such as (9), reflecting a priori ignorance about the extent to which observed cell means will be shrunk towards model-based prior means. In view of dependency of the weight on the prior mean in the shrinkage formula (9) on the within cell variances, some summary of the cell-specific first stage variances is required in order to define the uniform shrinkage prior for \(\Sigma\). We used the harmonic mean of the cell-specific variances, \(\bar{V} = \left(\frac{1}{J} \sum_{j=1}^{J} \sum_{i=1}^{C_j} V_{ij}^{-1}\right)^{-1}\) as the summary measure.

Adopting a uniform prior for the corresponding shrinkage parameter \(\bar{W} = C_j^{-1} \bar{V} (\Sigma + C_j^{-1} \bar{V})^{-1}\) implies \(p(\Sigma) \propto \|\bar{V} + \Sigma\|^{-(d+1)}\) (Daniels, 1999).
Rather than specifying a prior for the Dirichlet Process precision parameter, $\alpha$, in the applications described below we explore the impact on inferences and predictive disclosure risk of fixing $\alpha$, at different values. As $\alpha$, tends to infinity the model described by equations (10)-(13) reduces to a fully parametric hierarchical model whereas as $\alpha$ tends to zero the contribution of the parametric part of the model given by (12) and (13) declines and, in the limit, the model reduces to cell-specific Bayesian bootstrap sampling (Rubin, 1981; Lo 1988).
5 Fitting hierarchical and non-parametric Bayesian models

5.1 Fitting hierarchical Bayesian Poisson log-linear models

The fitting of hierarchical Bayesian Poisson log-linear models for cross-classified categorical data has been discussed elsewhere (Christiansen & Morris, 1997) and the PRIMM software for fast fitting of the model is freely available from Statlib (http://www.lib.stat.cmu.edu) as a S+ function (Christiansen & Morris, 1997). The model can, of course, also be fit using Bayesian modelling software, such as WinBugs (Spiegelhalter, 2003) and this allows a wider range of prior specifications for the hyper-parameters than is available in PRIMM.

5.2 Fitting the hierarchical non-parametric Bayesian model for numerical variables

5.2.1 Exploiting the Polya urn representation of the Dirichlet Process

The fitting of the hierarchical non-parametric model for numerical data is less standard than for the hierarchical Poisson log-linear models. The Dirichlet Process model is equivalent to an infinite mixture of point masses (Sethuraman 1994) and, as noted in section 4.5, finite mixture models can be used to approximate the infinite mixture (Ishwaran & James, 2001; Ohlssen et al, 2007). However, in the case of the model (10)-(13), a more direct and, in our experience faster, computational approach is available by exploiting the Polya urn data model implied by the Dirichlet Process model (Blackwell & McQueen, 1983).

The marginal model for the data, implied by (10) and (11) is a set of cell-specific Polya urn models, centered on multivariate normal models, with empty initial data urns. Under this model, given model parameters, the first observation in each cell, \( Y_j \), is assumed to be drawn from a multivariate normal model. Subsequent observations are assumed to be drawn either from the model or from the data-urn with probabilities which depend on the precision parameter \( \alpha \) and the size of the data-urn, as described in 4.4. Formally, for the purposes of inference for the parameters of the multivariate normal centering model we use the following model for the data, in place of (10) and (11).

\[
Y_{j,obs} | Y_j, \mu_j, \alpha \sim PU(\alpha, G_{\theta_j}, \emptyset), \quad j = 1, \ldots, J, \tag{14}
\]

where \( Y_j \) is the \( C_j \times d \) matrix of multivariate observations for the \( C_j \) people in cell \( j \), \( G_{\theta_j} \) denotes a \( MVN(\mu_j, V_j) \) distribution and the notation \( PU(\alpha, G, D) \) indicates Polya urn sampling model, with precision parameter, \( \alpha \), centering distribution \( G \) and initial data-urn \( D \), with \( \emptyset \), denoting an empty data urn. Note that under the Polya urn model the independence of individual observations is lost as the values of observations stored in the data-urn impact on the probability model for subsequent observations, since the model assigns positive probability to values stored in the urn. The loss of independence of individual observations is a result of integrating out the random distribution, \( F_j \).

Assuming (14), and a \( MVN(\mu_j, V_j) \) centering distribution, \( G_{\theta_j} \), the likelihood function for data \( Y_{obs} = (Y^obs_1, Y^obs_2, \ldots, Y^obs_J) \), given cell means \( \mu = (\mu_1, \mu_2, \ldots, \mu_J) \), cell-specific variances \( V = (V_1, \ldots, V_J) \) and Dirichlet Process parameter \( \alpha \), has the following form:

\[
p(Y_{obs} | \mu, V, \alpha) = \prod_j pu(Y^obs_j | \alpha, G_{\theta_j}, \emptyset) \tag{15}
\]

where the notation \( pu(Y_j | \alpha, G_{\theta_j}, D) \) denotes the probability function corresponding to Polya urn sampling for \( Y_j \), given Dirichlet Process precision parameter, \( \alpha \), centering distribution \( G_{\theta_j} \) and initial data urn \( D \), with \( D = \emptyset \) indicating an empty initial data urn. The contribution of the jth cell to the likelihood in (15) is therefore

\[
pu(Y^obs_j | \alpha, G_{\theta_j}, \emptyset) = \text{mvn}(Y_j, \mu_j, V_j) \prod_{i=1}^{C_j} \left( \frac{\alpha}{\alpha + i - 1} \right) \text{mvn}(Y^i_j, \mu_j, V_j) + \frac{1}{(\alpha + i - 1)} \sum_{i=1}^{C_j} \delta(Y^i_j, Y^obs_j) \tag{16}
\]
\[
\delta(Y_{ij}, Y_{i\cdot}) = 1, \text{ if } Y_{ij} = Y_{i\cdot}, \ \delta(Y_{ij}, Y_{\cdot j}) = 0, \text{ if } Y_{ij} \neq Y_{i\cdot}.
\]

The derivation of the likelihood contributions in (16) follows directly from the definition of the Polya urn sampling process. The first observation in each cell is drawn from the model and hence contributes \( \text{mvn}(Y_{ij} \mid \mu_j, V_j) \) to the likelihood. All subsequent observations may be drawn from the model (with probability \( \alpha / (\alpha + i - 1), i = 2 \ldots C_j \)) or from the data urn (with probability, \( (i - 1) / (\alpha + i - 1), i = 1 \ldots C_j \)). If an observed value of \( Y_{ij} \), is not represented in the (i-1) values in the cell-specific data urn, then it must have been generated from the model and therefore contributes \( (\alpha / (\alpha + i - 1)) \text{mvn}(Y_{ij} \mid \mu_j, V_j) \) to the likelihood. On the other hand if an observed value of \( Y_{ij} \) is represented among the preceding (i-1) values stored in the data urn, then it may have been drawn from the model or the data-urn. If the latter, then the probability \( Y_{ij} \) is equal to any particular data urn value is \( 1 / (i - 1) \). The contribution of a non-unique \( Y_{ij} \) to the likelihood is therefore.

\[
\frac{\alpha}{(\alpha + i - 1)} \text{mvn}(Y_{ij} \mid \mu_j, V_j) + \frac{(i - 1)}{(\alpha + i - 1)} \times \frac{1}{(i - 1)} \sum_{k=1}^{i-1} \delta(Y_{i\cdot j}, Y_{ik})
\]

\[
= \frac{\alpha}{(\alpha + i - 1)} \text{mvn}(Y_{ij} \mid \mu_j, V_j) + \frac{1}{(\alpha + i - 1)} \sum_{k=1}^{i-1} \delta(Y_{i\cdot j}, Y_{ik}).
\]

Compared to likelihoods derived from i.i.d. sampling from standard parametric models the Polya urn likelihood has some unusual features. The likelihood contribution of the \( i \)th observation in cell \( j \) depends on the values of the preceding \((i-1)\) observations in that cell, through the composition of the cell-specific data-urn, and the likelihood is, therefore, not simply the product of individual independent contributions as in the i.i.d. sampling situation. Moreover, the order in which the observations are recorded affects the contribution of each cell to the likelihood. In situations in which there is no natural ordering for the recording of observations this seems strange, but is nevertheless a feature of the Polya urn model and, therefore, a consequence of adopting the Dirichlet Process prior for the data distribution. For cells with large sample sizes \( (C_j) \), the order effects could be expected to be minor because the effects of the different possible permutations should approximately cancel out. For small cells the order in which observations are recorded could impact on the contribution of the cell to the likelihood, however the overall contribution of such cells to the likelihood is likely to be small.

Consequently, order effects are unlikely to make a major contribution to the likelihood.

### 5.2.2 A Gibbs sampling algorithm

With the Polya urn data model (14) replacing (10) and (11), parameters of the centering distribution given by \( \theta_j = (\mu_j, V_j) \) and \( V_j, \ j = 1 \ldots J \) and \( \alpha \) assumed fixed, the unknown parameters in the model are therefore \( (\mu_j, j = 1 \ldots J, \gamma, \Sigma) \). We assume \( p(\gamma, \Sigma) = p(\gamma) p(\Sigma) \), and consequently \( p(\mu, \gamma, \Sigma) = p(\mu \mid \gamma, \Sigma) p(\gamma) p(\Sigma) \) where \( \mu = (\mu_1, \ldots, \mu_J) \). Gibbs sampling provides a convenient method for simulating the posterior distribution for the unknown parameters. The Gibbs sampler alternates between sampling from the conditional posterior distribution of each of the unknowns, conditional on all other parameters. Denoting the observed data by \( Y^\text{obs} \) these conditional posterior distributions can be described as follows:

\[
p(\mu \mid \gamma, \Sigma, X, V, \alpha, Y^\text{obs}) \propto p(Y^\text{obs} \mid \mu, V, \alpha) p(\mu \mid \gamma, \Sigma)
\]

\[
= \prod_j p(Y^\text{obs}_{ij} \mid \mu_j, V_j, \alpha) p(\mu_j \mid X_j, \gamma, \Sigma)
\]

\[
p(\gamma \mid \mu, \Sigma, X, V, \alpha, Y^\text{obs}) \propto p(\mu \mid \gamma, \Sigma, X) p(\gamma)
\]

\[
p(\Sigma \mid \mu, \gamma, X, V, \alpha, Y^\text{obs}) \propto p(\mu \mid \Sigma, X, \gamma) p(\Sigma)
\]

where \( X = (X_1, \ldots X_J) \); \( V = (V_1, \ldots V_J) \); \( Y^\text{obs}_{ij} = (Y_{ij_1}, \ldots Y_{ij_{C_j}}) \); \( Y^\text{obs} = (Y^\text{obs}_{ij_1}, \ldots Y^\text{obs}_{ij_{C_j}}) \).
The full conditionals (18) and (19) are identical to those that pertain under a fully parametric hierarchical multivariate normal model (Graham, 2005) and it is only the full conditional for the cell-specific means (17) which is complicated by moving from a standard parametric formulation of the model for the data to a Dirichlet Process model. Under a normal or uniform prior for the hyper-regression parameters, $\gamma$, standard normal theory results (Lindley & Smith, 1972) show that the full conditional distribution for the regression hyper-parameters in (18) is multivariate normal with mean

$$\gamma_{\text{post}} = \Omega_{\text{post}} (\sum_j Z_j \Sigma^{-1} \mu_j + \Omega^{-1} \gamma_{\text{prior}})$$

and variance

$$\Omega_{\text{post}} = (\sum_j Z_j \Sigma^{-1} Z_j + \Omega^{-1})^{-1}$$

Consequently the full conditional distribution for hyper-regression parameters is easily simulated. Under a uniform prior for the regression hyper-parameters, the full-conditional for these parameters reduce to a multivariate normal with mean $\gamma_{\text{post}} = \Omega_{\text{post}} (\sum_j Z_j \Sigma^{-1} \mu_j)$ and variance $\Omega_{\text{post}} = (\sum_j Z_j \Sigma^{-1} Z_j)^{-1}$.

Under the multivariate uniform shrinkage prior the full conditional for the variance hyper-parameter cannot be simulated directly, however a posterior sample from (19), can be obtained by using a Metropolis-Hastings step to correct draws from an Inverse-Wishart distribution (Graham, 2005)\(^1\). Direct sampling from the full conditional for the cell means, given by (17) does not appear possible. However the full conditional for the $\mu_j, j = 1, \ldots, J$ that would pertain under a multivariate normal model for the numerical data $Y_{ij} \sim MVN(\mu_j, V_j), \ i = 1, \ldots, C_j, j = 1, \ldots, J$ is the product over $j$ of multivariate normal densities with mean given by (9) and variance

$$V(\mu_j | Y^{\text{obs}}, X_j, \gamma, \Sigma) = C_j^{-1} V_j (C_j^{-1} V_j + \Sigma)^{-1} \Sigma \text{ (Lindley and Smith, 1972).}$$

It is therefore straightforward to sample from this approximation to the full conditional or from an over-dispersed variant. This can serve as an approximation which can be corrected by a Metropolis-Hastings step.

To formalise the approach let the approximate full conditional based on a fully parametric hierarchical multivariate normal be denoted

$$p_{\text{approx}}(\mu | Y^{\text{obs}}, X, V, \Sigma) = \prod_j \text{mvn}(\mu_j | \mu_{\text{post}}^j, \psi_j)$$

(20)

where $\mu_{\text{post}}^j = W_j Z_j \gamma + (I - W_j) Y_j, \ W_j = C_j^{-1} V_j (C_j^{-1} V_j + \Sigma)^{-1}$, $\psi_j = W_j \Sigma$. In view of the multivariate normal data model underlying the approximate full conditional, this distribution can also be written as follows

\(^1\) There is a mistake in the presentation of the Metropolis-Hastings acceptance probabilities in final paragraph of appendix A.3 of Graham (2005). The correct acceptance probabilities are min(1, $p(\Sigma^{(i)}) / p(\Sigma^{(i-1)})$) where the t superscript refers to the $t^{\text{th}}$ iteration of the Gibbs sampler.
Confidentialising Microdata

\[ p_{\text{approx}}(\mu | Y, X, V, \gamma, \Sigma) = \left( \prod_j \prod_{i=1}^{c_j} \text{mvn}(Y_{ij} | \mu_j, V_j) \right) \left( \prod_j \text{mvn}(\mu_j | Z_j, \gamma, \Sigma) \right) \]

whereas using the Polya urn characterisation of the Dirichlet Process and hence the Polya urn data model, the true full conditional can be written

\[ p(\mu | Y^{\text{obs}}, X, V, \gamma, \Sigma) = \prod_j p(u(Y_j | \alpha, G_{\theta_j}, \emptyset) \times \text{mvn}(\mu_j | Z_j, \gamma, \Sigma) \]

where \( \theta_j = (\mu_j, V_j) \) and \( G_{\theta_j} \) denotes a multivariate normal distribution with mean \( \mu_j \) and variance \( V_j \). Consequently the ratio of the true full conditional to the approximation based on the full parametric hierarchical multivariate normal model is

\[ w = \prod_j \frac{p(u(Y_j | \alpha, G_{\theta_j}, \emptyset) \times \text{mvn}(\mu_j | Z_j, \gamma, \Sigma)}}{\text{mvn}(Y_j | \mu_j, C_j^{-1}V_j)} \]

Therefore, a Metropolis-Hastings procedure for sampling from the full conditional in (17) at the \( t \)th iteration of the Gibbs Sampler can proceed by completing the following steps:

(I) Generate \( \mu_j^{\text{prop}} \), for \( j = 1 \ldots J \) from (20), with \( \gamma \) and \( \Sigma \), set at their most recently generated value,

(II) For each \( j \), compute the ratio of importance sampling ratios

\[ r_j^{(t)} = \frac{p(u(Y_j | \alpha, G_{\theta_j}^{\text{prop}}, \emptyset) \times \text{mvn}(\mu_j | Z_j, \gamma, \Sigma))}{\text{mvn}(Y_j | \mu_j, C_j^{-1}V_j)} \]

where \( G_{\theta_j}^{\text{prop}} \) denotes a multivariate normal distribution with mean and variance given by

\[ \theta_j = (\mu_j^{\text{prop}}, V_j), G_{\theta_j}^{\text{prop}} \]

denotes a multivariate normal distribution with parameter vector \( \theta_j = (\mu_j^{(t-1)}, V_j) \), and \( \mu_j^{(t-1)} \) is the value of \( \mu_j \), sampled at the (t-1)th iteration of the Gibbs sampler.

(III) For \( j = 1 \ldots J \), accept \( \mu_j^{\text{prop}} \) with probability \( \min(1, r_j^{(t)}) \). If a generated \( \mu_j^{\text{prop}} \) value is accepted set \( \mu_j^{(t)} = \mu_j^{\text{prop}} \), otherwise set \( \mu_j^{(t)} = \mu_j^{(t-1)} \).

Note that since the Polya urn likelihood places less weight on the parametric model than does the fully parametric model likelihood it is likely that the full conditional (17) based on the Polya urn model will be more dispersed than the fully parametric conditional distribution \( p_{\text{approx}}(\mu | Y^{\text{obs}}, X, V, \Sigma) \), given in (20). Consequently in order to increase the chance that the proposal distribution \( p_{\text{approx}}(\mu | Y^{\text{obs}}, X, V, \Sigma) \) covers the full conditional it may be prudent to modify

\[ p_{\text{approx}}(\mu | Y^{\text{obs}}, X, V, \Sigma) \]

by inflating the cell-specific variance terms, for example by reducing the cell size used in the calculation of the variance of the mean by some suitable fraction. For some \( \kappa \) in (0,1) this would mean substituting \( \kappa C_j^{-1}V_j \) for \( C_j^{-1}V_j \), in the variance and shrinkage formulae immediately following (20) and in subsequent formulae for importance sampling ratios.
6 Generating multiply imputed synthetic datasets from the combined categorical and numerical data models

Having fitted the imputation model the final step in the production of multiply imputed synthetic data is to generate multiple draws from the posterior predictive distribution for new data, given the observed data. Under standard parametric imputation models this involves first drawing from the posterior distribution of the model parameters and then generating new data from the data model with parameters set to the drawn values. In the case of Dirichlet-Process model for the data, there is, in principle, an additional step inserted between the above two steps which involves drawing distribution functions from the posterior over the space of distribution. However, in practice, because of the Polya urn representation of the Dirichlet Process this additional step can be left implicit, with data being drawn from a Polya urn model, conditionally on the Dirichlet Process precision parameter and parameters of the centering model.

To simplify notation let $\psi$ denote all parameters required for specification of the numerical and categorical data models, $F_Y = (F_{Y_1}, \ldots, F_{Y_J})$ denote the collection of cell-specific distributions for the numerical variable and let $G$ denote the form of the centering distribution for the Dirichlet Process models, with full specification of the latter depending on elements of $\psi$, denoted $s_j(\psi)$, for the $j^{th}$ cell-specific Dirichlet Process model. The posterior distribution for new data, comprising cell counts, $C_{new} = (C_{new,1}, \ldots, C_{new,J})$ and numerical data $Y_{new} = (Y_{new,1}, \ldots, Y_{new,J})$, given the observed data can be represented as follows:

$$p(C_{new}, Y_{new} | C_{obs}, Y_{obs}, G) = \int \int \int p(C_{new}, Y_{new} | C_{obs}, Y_{obs}, F_Y, \psi, G) p(F_Y | C_{obs}, Y_{obs}, \psi, G) \times$$

$$p(\psi | C_{obs}, Y_{obs}, G) dF_Y d\psi$$

$$= \int p(C_{new} | C_{obs}, \psi) p(\psi | C_{obs}, Y_{obs}, G) \int p(Y_{new} | C_{new, obs}, Y_{obs, obs}, F_Y, \psi, G) p(F_Y | C_{obs}, Y_{obs, obs}, \psi, G) dF_Y d\psi$$

The inner integration, over the posterior for the distribution of the numerical variable results in the Polya urn model for the data, centred on $G$, with initial data urns defined by the observed data (Blackwell & McQueen, 1973). Under the cell-specific imputation model for numerical data given by (10) to (13) this integration results in cell-specific Polya-urn models and the posterior predictive distribution for new data can be written

$$p(C_{new}, Y_{new} | C_{obs}, Y_{obs}, G) = \prod_{j=1}^{J} \left[ p_u(Y_{new, j} | \alpha, G_{s_j(\psi), Y_{obs, j}}, C_{new, j}) \right] \times$$

$$p(\psi | C_{obs, obs}, Y_{obs, obs}, G) \prod_{j=1}^{J} \int p(C_{new, j} | \psi) p(\psi | C_{obs, obs}, Y_{obs, obs}, G) d\psi$$

$$= \prod_{j=1}^{J} p_u(Y_{new, j} | \alpha, G_{s_j(\psi), Y_{obs, j}}) \times$$

$$p(C_{new, j} | \psi) p(\psi | C_{obs, obs}, Y_{obs, obs}, G) d\psi$$

(21)

A Monte Carlo approach to evaluating the integral in (21) involves repeatedly generating from the posterior distribution of the model parameters, then generating new cell counts, $C_{new, j}$, from the posterior predictive distribution of the cell counts, given the generated categorical data model parameters, and finally, for each cell, $j = 1, \ldots, J$ drawing $C_{new, j}$ numerical data records from the cell-specific Polya urn, with initial data-urn set to $Y_{obs, j}$ and parameters of the centering model set to the relevant component of the generated parameter vector, $\psi$. This Monte Carlo evaluation of (21) generates multiply imputed synthetic datasets.

More specifically, given the hierarchical Poisson log-linear model for categorical data described in section 3, the non-parametric hierarchical Bayesian model for numerical data outlined in section 4, the multiply imputed synthetic datasets can be generated using the six steps given below. If posterior
samples for the first-stage parameters (Poisson parameters, \( \lambda_j \), and cell-means \( \mu_j \)) have been stored then steps two and five can be omitted. For data structures with large number of cells, storage may be an issue for the posterior sample of first-stage parameters (Poisson parameters and cell-specific means) and consequently, though computationally inefficient, it may be judged preferable to re-generate these parameters in the course of drawing from the posterior predictive distribution.

Draw from the posterior distribution of the hyper-parameters from the hierarchical Poisson log-linear model, ie draw

\[
\beta^*, \xi^* \text{ from } p(\beta, \xi | C^{obs}).
\]

Typically this step will be implemented by drawing from the stored posterior sample of the hyper-parameters of the categorical data model.

Draw expected cell counts (Poisson cell parameters) given the hyper-parameters drawn in step 1, ie draw \( \lambda_j^{\text{indep}} \sim \lambda_j | \beta^*, \xi^*, X_j, C_j^{obs}, j = 1 \ldots J \), which under the hierarchical Poisson log-linear model is equivalent to

\[
\lambda_j^{\text{indep}} \sim \text{GAMMA}(\xi^* + C_j, \xi / \eta_j + 1), \quad \eta_j = \exp(X_j \beta^*), \quad j = 1 \ldots J
\]

Given the generated expected cell counts draw synthetic cell counts from independent Poisson distributions, ie

\[
C_j^{\text{syn}} | \lambda_j^{\text{indep}} \sim \text{POISSON}(\lambda_j^*), \quad j = 1 \ldots J
\]

Draw from the posterior distribution for the hyper-parameters of the hierarchical multivariate normal model which serves as the centering distribution of the Dirichlet Process prior, ie draw \( \gamma^*, \Sigma^* \) from \( p(\gamma, \Sigma | Y^{obs}, X, \alpha) \).

This step could be implemented by drawing from a stored sample from the posterior of the hyper-parameters obtained from a Gibbs sampling algorithm, as described in section 3.

For \( j = 1 \ldots J \) draw cell means, \( \mu_j^* \) from

\[
\mu_j^* \text{ from } p(\mu_j | Y_j^{obs}, \gamma^*, \Sigma, X_j, V_j, \alpha)
\]

where \( p(\mu | Y_j^{obs}, \gamma^*, \Sigma, X_j, V_j) \propto \text{pu}(Y_j^{obs} | \alpha, G_{\theta_j} \emptyset) \times \text{mvn}(\mu_j | Z_j, \gamma, \Sigma) \)

and \( G_{\theta_j} \) denotes a multivariate normal distribution with mean and variance \( \theta_j = (\mu_j, V_j) \).

(VI) For \( j = 1 \ldots J; \) for \( i = 1 \ldots C_j^{\text{syn}} \), draw \( Y_{ij}^{\text{syn}} \) from

\[
Y_{ij}^{\text{syn}} \sim \text{PU}(\alpha, G_{\theta_j}^*, Y_j^{obs})
\]

where \( G_{\theta_j}^* \) denotes a multivariate normal distribution with mean and variance \( \theta_j^* = (\mu_j^*, V_j) \).

Repeating the six steps above \( M \) times, yields \( M \) synthetic sets of cell counts and numerical data for each cell. Synthetic datasets in the usual unit-record format can be obtained by converting each synthetic cell count, \( C_j^{\text{syn}} \), (generated in step 3, above) to a matrix with \( C_j^{\text{syn}} \) rows, each equal to the cell descriptor \( X_j \) and concatenating this matrix with the \( C_j^{\text{syn}} \times d \) matrix of numerical values generated in step 6.
If a posterior sample for the cell means, \( \mu_j, j = 1, \ldots, J \) has not been saved in the course of fitting the model then the \( \mu_j \) need to be regenerated in step 5. The Metropolis-Hastings algorithm for drawing from the conditional posterior distribution as part of the Gibbs sampler described in section 4 is unlikely to be an attractive option for generating the \( \mu_j \) in the above algorithm for generating synthetic data, because the Metropolis – Hastings algorithm would need to be run to convergence for each cell, for each of the M synthetic datasets\(^2\). An alternative approach which better-suited for the imputation phase of the computations is sampling importance resampling (Gelman et al, 2004, p. 316-317). For each synthetic dataset and for each cell this involves generating \( K > 1 \) values of \( \mu_j \), denoted \( \mu_j^{(1)}, \ldots, \mu_j^{(K)} \), from the multivariate normal approximation given in (20), computing the importance sampling weights

\[
v_{jk} = \frac{pu(Y_j; G_{\theta_j}^{(k)}, \emptyset)}{mvn(\bar{Y}_j; \mu_j^{(k)}, C_j^{-1}V_j)}, \quad k = 1, \ldots, K,
\]

where \( G_{\theta_j}^{(k)} \) is MVN(\( \mu_j^{(k)}, V_j \)) and re-sampling the \( K \) candidate \( \mu_j \) values with probability proportional to the importance sampling weights to yield a single \( \mu_j \).

\(^2\) In the model fitting phase, the Metropolis-Hastings step for generating the first stage means is embedded with a Gibbs sampling algorithm. Convergence issues pertain to the Gibbs-Sampler as a whole and this means that the Metropolis-Hastings algorithm does not need to be run to convergence at each iteration of the Gibbs sampler.
7 Empirical comparison of parametric and non-parametric hierarchical models

7.1 Data source and methods

7.1.1 Data

We based our comparison of the performance of fully parametric and non-parametric hierarchical Bayesian imputation models for numerical data on a subset of the confidentialised unit record file of the 2003 Household Income survey (IS2003) conducted by Statistics New Zealand. For this initial comparison of imputation models for numerical data we treat the IS2003 CURF (henceforth just CURF) as real data, fit both fully parametric and non-parametric hierarchical Bayesian imputation models and use the models to generate synthetic versions of the CURF. We then compare the inferences obtained from the CURF with those obtained from the synthetic versions of the CURF and also, compute basic diagnostics of confidentiality risk.

Details of the construction of the CURF and basic information about the Household Income Survey are provided in a Statistics New Zealand report (Statistics New Zealand, 2005). In brief, the Household Income Survey is run as a supplement to the Household Labour Force Survey (HLFS) and collects information on income from all sources, in addition to employment data collected as part of the HLFS. To construct the CURF, Statistics New Zealand top-code and sometimes bottom-code numerical variables such as income and hours worked and collapse categories of some categorical variables. Some records are also deleted from the CURF. The HLFS uses a complex sample design and replicate weights are included in the CURF to facilitate accounting for the complex design in analyses. However, for the exercise reported here we have treated the CURF as a simple random sample in order focus the presentation on the main issue of the relative performance of parametric and non-parametric hierarchical Bayesian imputation models.

We extracted a subset of the CURF using the following inclusion criteria: positive average weekly income, in paid work at least one hour per week, aged between 25 and 64. In addition to average weekly income, hours worked and age (in five year categories) we also used the variables sex, ethnic group (European, Māori, Pacific, Other, Māori combinations, Other combinations) and educational achievement (no qualifications, secondary school qualifications, trade qualifications, university qualifications, other post-secondary qualifications) in the analyses and comparisons reported here. The motivation for the inclusion criteria and selection of variables for CURF subset was to produce a dataset that could be used for analysis of the effect of education on income and variation in this relationship by sex and ethnicity. However we emphasise that the analyses reported below were not focussed on this issue, per se, but on investigating the quality of synthetic versions of the CURF data.

We compared the synthetic data inferences obtained under eight imputations models with those obtained from the CURF (treating the CURF as real data) for both descriptive statistics (means, medians, quartiles) and parameters of linear and generalised linear models. We used variety of metrics for these comparisons, including absolute differences in point estimates, relative credible interval lengths, comparing synthetic and real data intervals, and a measure of data utility proposed by Karr et al (2006) based on credible interval overlap. The latter measure is the average of the fraction of the posterior distribution for a parameter obtained from the real data that falls within the central 95% credible interval obtained from synthetic data, and the fraction of the synthetic data posterior distribution which falls within the credible interval obtained from real data. Identical credible intervals result in a data-utility index of 0.95. This measure of data-utility is specific to particular inference targets, such as individual model parameters rather than comparing joint distributions of real and synthetic data in a global manner.

Some basic descriptive information on the CURF subset is given in Tables 1 and 2, which summarise the distribution of weekly income and hours worked for the CURF subset and for 35 sub-groups used subsequently in the evaluation of the performance of the synthetic data methods in estimating descriptive statistics for a range of sample sizes. Columns two and three in Table 1 contrast actual and synthetic sample sizes for the sub-groups. The reported synthetic cell sizes were obtained as the average of the sample sizes obtained from 100 imputations under a hierarchical Poisson log-linear imputation model described in 7.1.2, below. The synthetic and actual sample sizes are in close agreement. The focus of this report is on the performance of synthetic data methods for numerical data so we do not comment further on the performance of the imputation model for categorical data.
Table 1

<table>
<thead>
<tr>
<th>Group</th>
<th>Sample size</th>
<th>Synthetic(^1) size</th>
<th>Lower quartile</th>
<th>Median</th>
<th>Upper quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>11267</td>
<td>11267</td>
<td>440</td>
<td>630</td>
<td>860</td>
</tr>
<tr>
<td>females</td>
<td>5739</td>
<td>5742.5</td>
<td>440</td>
<td>520</td>
<td>730</td>
</tr>
<tr>
<td>males</td>
<td>5529</td>
<td>5524.5</td>
<td>450</td>
<td>650</td>
<td>900</td>
</tr>
<tr>
<td>European</td>
<td>8665</td>
<td>8667</td>
<td>450</td>
<td>650</td>
<td>900</td>
</tr>
<tr>
<td>Māori</td>
<td>1015</td>
<td>1005</td>
<td>420</td>
<td>580</td>
<td>790</td>
</tr>
<tr>
<td>Pacific</td>
<td>541</td>
<td>547.5</td>
<td>420</td>
<td>520</td>
<td>650</td>
</tr>
<tr>
<td>other ethnicity</td>
<td>643</td>
<td>638</td>
<td>430</td>
<td>630</td>
<td>850</td>
</tr>
<tr>
<td>Māori combinations</td>
<td>346</td>
<td>346.5</td>
<td>440</td>
<td>620</td>
<td>817.5</td>
</tr>
<tr>
<td>Other combinations</td>
<td>57</td>
<td>58</td>
<td>330</td>
<td>560</td>
<td>720</td>
</tr>
<tr>
<td>No quals</td>
<td>2140</td>
<td>2137.5</td>
<td>360</td>
<td>510</td>
<td>660</td>
</tr>
<tr>
<td>secondary school quals</td>
<td>2315</td>
<td>2318.5</td>
<td>400</td>
<td>580</td>
<td>770</td>
</tr>
<tr>
<td>trade quals</td>
<td>4235</td>
<td>4223.5</td>
<td>480</td>
<td>680</td>
<td>880</td>
</tr>
<tr>
<td>university quals</td>
<td>1749</td>
<td>1736</td>
<td>620</td>
<td>870</td>
<td>1200</td>
</tr>
<tr>
<td>Post-secondary quals</td>
<td>828</td>
<td>833.5</td>
<td>450</td>
<td>640</td>
<td>832.5</td>
</tr>
<tr>
<td>female, European</td>
<td>4420</td>
<td>4427.5</td>
<td>340</td>
<td>530</td>
<td>730</td>
</tr>
<tr>
<td>female, Māori</td>
<td>540</td>
<td>539.5</td>
<td>370</td>
<td>515</td>
<td>700</td>
</tr>
<tr>
<td>female, Pacific</td>
<td>251</td>
<td>252</td>
<td>400</td>
<td>480</td>
<td>630</td>
</tr>
<tr>
<td>female, other ethnicity</td>
<td>328</td>
<td>325.5</td>
<td>340</td>
<td>560</td>
<td>770</td>
</tr>
<tr>
<td>female, Māori combinations</td>
<td>165</td>
<td>166</td>
<td>390</td>
<td>520</td>
<td>700</td>
</tr>
<tr>
<td>female, other combinations.</td>
<td>35</td>
<td>35.5</td>
<td>365</td>
<td>530</td>
<td>680</td>
</tr>
<tr>
<td>male, European</td>
<td>4245</td>
<td>4240.5</td>
<td>600</td>
<td>770</td>
<td>1040</td>
</tr>
<tr>
<td>male, Māori</td>
<td>475</td>
<td>469.5</td>
<td>520</td>
<td>670</td>
<td>880</td>
</tr>
<tr>
<td>male, Pacific</td>
<td>290</td>
<td>291</td>
<td>470</td>
<td>560</td>
<td>670</td>
</tr>
<tr>
<td>male, other ethnicity</td>
<td>315</td>
<td>316</td>
<td>515</td>
<td>580</td>
<td>940</td>
</tr>
<tr>
<td>male, Māori combinations</td>
<td>181</td>
<td>180.5</td>
<td>540</td>
<td>700</td>
<td>960</td>
</tr>
<tr>
<td>male, other combinations.</td>
<td>22</td>
<td>22</td>
<td>322.5</td>
<td>570</td>
<td>742.5</td>
</tr>
<tr>
<td>female, no quals</td>
<td>1075</td>
<td>1077</td>
<td>280</td>
<td>400</td>
<td>540</td>
</tr>
<tr>
<td>female, secondary quals</td>
<td>1352</td>
<td>1360.5</td>
<td>330</td>
<td>500</td>
<td>660</td>
</tr>
<tr>
<td>female, trade quals</td>
<td>2024</td>
<td>2013.5</td>
<td>380</td>
<td>550</td>
<td>740</td>
</tr>
<tr>
<td>female university quals</td>
<td>911</td>
<td>900</td>
<td>530</td>
<td>770</td>
<td>1030</td>
</tr>
<tr>
<td>female, post-secondary quals</td>
<td>377</td>
<td>377.5</td>
<td>330</td>
<td>490</td>
<td>680</td>
</tr>
<tr>
<td>male, no quals</td>
<td>1065</td>
<td>1071</td>
<td>480</td>
<td>600</td>
<td>770</td>
</tr>
<tr>
<td>male, secondary quals</td>
<td>963</td>
<td>961.5</td>
<td>520</td>
<td>680</td>
<td>910</td>
</tr>
<tr>
<td>male, trade quals</td>
<td>2211</td>
<td>2204.5</td>
<td>610</td>
<td>770</td>
<td>1000</td>
</tr>
<tr>
<td>male university quals</td>
<td>838</td>
<td>836</td>
<td>720</td>
<td>1040</td>
<td>1500</td>
</tr>
<tr>
<td>male, post-secondary quals</td>
<td>451</td>
<td>455</td>
<td>580</td>
<td>750</td>
<td>970</td>
</tr>
</tbody>
</table>

\(^1\)Average cell size for 100 synthetic datasets generated from a hierarchical Bayes imputation model fitted to the CURF age by sex by ethnicity by educational qualifications data.
Table 2

Quartiles for hours worked per week for the overall CURF dataset and for 36 subgroups

<table>
<thead>
<tr>
<th>Group</th>
<th>Lower quartile</th>
<th>Median</th>
<th>Upper quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>30</td>
<td>40</td>
<td>44.9</td>
</tr>
<tr>
<td>females</td>
<td>21</td>
<td>37</td>
<td>40</td>
</tr>
<tr>
<td>males</td>
<td>40</td>
<td>40</td>
<td>48</td>
</tr>
<tr>
<td>European</td>
<td>30</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td>Māori</td>
<td>32</td>
<td>40</td>
<td>44</td>
</tr>
<tr>
<td>Pacific</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>other ethnicity</td>
<td>33</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Māori combinations</td>
<td>32</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td>Other combinations</td>
<td>20</td>
<td>40</td>
<td>43.7</td>
</tr>
<tr>
<td>No quals</td>
<td>26</td>
<td>40</td>
<td>43</td>
</tr>
<tr>
<td>secondary school quals</td>
<td>29</td>
<td>40</td>
<td>40.5</td>
</tr>
<tr>
<td>trade quals</td>
<td>32.5</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td>university quals</td>
<td>35.9</td>
<td>40</td>
<td>46</td>
</tr>
<tr>
<td>Post-secondary quals</td>
<td>32</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td>female, European</td>
<td>20</td>
<td>36</td>
<td>40</td>
</tr>
<tr>
<td>female, Māori</td>
<td>24</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>female, Pacific</td>
<td>30.75</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>female, other ethnicity</td>
<td>27</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>female, Māori combinations</td>
<td>24</td>
<td>37.5</td>
<td>40</td>
</tr>
<tr>
<td>female, other combinations</td>
<td>20</td>
<td>37.5</td>
<td>45</td>
</tr>
<tr>
<td>male, European</td>
<td>40</td>
<td>40.5</td>
<td>48.3</td>
</tr>
<tr>
<td>male, Māori</td>
<td>40</td>
<td>40</td>
<td>46.5</td>
</tr>
<tr>
<td>male, Pacific</td>
<td>40</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td>male, other ethnicity</td>
<td>38.7</td>
<td>40</td>
<td>44.25</td>
</tr>
<tr>
<td>male, Māori combinations</td>
<td>40</td>
<td>40</td>
<td>48</td>
</tr>
<tr>
<td>male, other combinations</td>
<td>20.25</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>female, no quals</td>
<td>20</td>
<td>32</td>
<td>40</td>
</tr>
<tr>
<td>female, secondary quals</td>
<td>21</td>
<td>36.8</td>
<td>40</td>
</tr>
<tr>
<td>female, trade quals</td>
<td>23</td>
<td>37</td>
<td>40</td>
</tr>
<tr>
<td>female university quals</td>
<td>29</td>
<td>40</td>
<td>42</td>
</tr>
<tr>
<td>female, post-secondary quals</td>
<td>20</td>
<td>36.8</td>
<td>40</td>
</tr>
<tr>
<td>male, no quals</td>
<td>40</td>
<td>40</td>
<td>48</td>
</tr>
<tr>
<td>male, secondary quals</td>
<td>40</td>
<td>40</td>
<td>48</td>
</tr>
<tr>
<td>male, trade quals</td>
<td>40</td>
<td>40</td>
<td>48</td>
</tr>
<tr>
<td>male university quals</td>
<td>40</td>
<td>41</td>
<td>50</td>
</tr>
<tr>
<td>male, post-secondary quals</td>
<td>40</td>
<td>42</td>
<td>50</td>
</tr>
</tbody>
</table>

1 Subgroup sample sizes are as given in Table 1

7.1.2 Imputation models

Results are presented below for eight synthetic data methods corresponding to eight different imputation models. The models all used a hierarchical Poisson log-linear model for the categorical data, but differed with respect to the model for the conditional distribution of numerical variables given the categorical variables. The hierarchical Poisson log-linear model included two-way interactions between the categorical variables with age represented as a quadratic spline with a single knot at age 45 (Greenland, 1995). The quadratic spline representation of age involves three components; a linear term, a quadratic term and a term defined to be zero for ages less than 45 but equal to the squared difference between age and 45 for ages 45 or older. Age categories were represented by their category midpoint.

The eight approaches to modelling the conditional distribution of income and working hours given the categorical predictors were defined by the scale on which income and working hours were modelled and the structure adopted for modelling the conditional distribution. We considered both models for weekly income and hours worked, specified on their original scale (ie dollars and hours) and, also models specified on a log-transformed scale. The four model structures adopted for each measurement scale (original and log-transformed) were a fully parametric hierarchical multivariate normal model, two non-parametric hierarchical imputation models based on a Dirichlet Process model with a hierarchical multivariate normal centering model and precision parameters $\alpha = 19$ and $\alpha = 9$, respectively, and a finite population Bayesian bootstrap, (Rubin, 1981; Lo, 1988) implemented separately for each cell defined by the categorical variables.
Note that the fully parametric hierarchical model corresponds to setting the precision parameter, $\alpha$, to infinity in a Dirichlet Process model centered on the hierarchical multivariate normal model, whereas the Bayesian bootstrap corresponds to $\alpha = 0$. In principle this means that the Bayesian bootstrap does not depend on an imputation model structure and always samples from the observed data. However our imputation of numerical variables is conditional on categorical variables represented as a cross-classified table and synthetic cell counts for cells which are empty in the real data can sometimes be positive. In these cases there is no data-urn for the Bayesian bootstrap to draw from and we therefore used draws from the fully parametric hierarchical imputation model to generate numerical values in such cases. Since synthetic cell counts for truly empty cells are likely to be small the overall impact of these model generated values on the synthetic dataset is likely to be small. In comparisons in which in which the fully parametric and non-parametric hierarchical imputation models were specified on the original scale we used the original scale fully parametric hierarchical model to supplement the Bayesian bootstrap for cells with positive synthetic cell counts but zero counts in the real data. Similarly we used the log scale fully parametric hierarchical imputation model to supplement the Bayesian bootstrap for comparisons in which the fully parametric and non-parametric hierarchical models were specified on the log scale.

The Bayesian bootstrap approach is not a realistic imputation method for synthetic data since it always (with the exception of empty cells) samples records from the real data and therefore provides little disclosure protection. However, since the Bayesian bootstrap corresponds to the extreme value of the Dirichlet Process precision parameter which places full weight on the data and essentially no weight on the parametric centering model, it provides a useful point of comparison, for evaluating methods which place some weight on a parametric model. In a sense, the Bayesian bootstrap represents the closest synthetic data can get to reproducing the inferences obtained from real data, since the Bayesian bootstrap re-samples from the data. Note, however, that because of Monte Carlo variation in the frequency with which real data records are sampled, the Bayesian bootstrap synthetic datasets do not exactly match the originally collected data.

The prior multivariate normal regression model for the hierarchical models included main effects and all two-way interactions between categorical predictors (age, sex, ethnic group and educational achievement). However, while the main effect for age was represented by a quadratic spline, as described above for the categorical data imputation model, only the linear component of the age effect was included in age-related interaction terms. This model simplification was adopted to simplify model-fitting and also to provide a simple means of investigating the effect of a mild form of prior imputation model mis-specification by considering analysis models which included the full quadratic spline representation of age in interaction terms.

### 7.1.3 Analysis models

We evaluated the performance of the synthetic data for three sets of 9 regression models corresponding to linear regression, linear regression on the log-scale and generalised linear modelling assuming Gamma distributed responses (weekly income and hours worked) and a log-link function. The models varied in complexity from models including only main effect terms to models including all two-way interactions between predictors. Regressions for the income variable included working hours as a predictor and separate models were fitted including working hours and log-transformed working hours. Since income is usually modelled on a log-scale (Heckman & Polachek, 1974; Hyslop, 2000; Maani, 2000, 2002) and to limit the volume of results presented only results for the log-linear regressions for weekly income are presented below. We present results for only five of the nine analysis models as these results suffice to illustrate the performance of the synthetic data methods. A set of spreadsheets containing results for all nine models for the linear regressions, log-linear regressions and Gamma generalised linear models is available from the author.

Some details of the five model structures for which results are presented are given in Table 3.
Table 3

Description of analysis models

<table>
<thead>
<tr>
<th>Model</th>
<th>Outcome variable</th>
<th>Predictors</th>
<th>Number of parameters</th>
<th>Relationship to prior imputation model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Original scale</td>
</tr>
<tr>
<td>4</td>
<td>Log weekly income</td>
<td>Age, sex, ethnicity, education, log(hours worked)</td>
<td>15</td>
<td>uncongenial</td>
</tr>
<tr>
<td>5</td>
<td>Log weekly income</td>
<td>Age, sex, ethnicity, education, log(hours worked), linear age by sex interaction</td>
<td>16</td>
<td>uncongenial</td>
</tr>
<tr>
<td>6</td>
<td>Log weekly income</td>
<td>Main effects as per model 4, plus full age by sex, ethnicity by qualifications, sex by qualifications, log(hours) by ethnicity and log working hours by sex interactions</td>
<td>48</td>
<td>uncongenial</td>
</tr>
<tr>
<td>7</td>
<td>Log weekly income</td>
<td>Age, sex, ethnicity, education, hours worked</td>
<td>15</td>
<td>uncongenial</td>
</tr>
<tr>
<td>9</td>
<td>Log weekly income</td>
<td>Main effects as per model 7, plus age by sex, ethnicity by qualifications, sex by qualifications, hours by ethnicity and hours by sex interactions</td>
<td>48</td>
<td>uncongenial</td>
</tr>
</tbody>
</table>

In all analysis models age was represented by a quadratic spline with a single knot at age 45, with each five year age group assigned its category mid-point. This representation of the age-group variable was used in both main effect and interaction terms, except for model 5 which included only the linear age term in an age by sex interaction.

The last two columns describe the relationship of analysis models to imputation models in terms of the congeniality of the analysis and imputation models (Meng, 1994) and we use this concept to structure the presentation of results. Congeniality of the analysis and imputation models refers to the degree of compatibility of the models, although in our case because the imputation models are hierarchical we interpret congeniality in terms of the relationship between the analysis model and the structure of prior models. We define the prior imputation model to be the regression structure of the multivariate-normal adopted for cell means given in (8). For the non-parametric imputation models this regression structure for cell means is incorporated into the hierarchical model which centers non-parametric data model. Multivariate normal models implicitly impose an assumption of linearity for the association between the components of the multivariate normal and this also needs to be taken into account in determining the congeniality of analysis and prior imputation models.

We consider analysis and imputation models to be congenial if they are specified on the same scale and the predictor terms included in the analysis model are a subset of the predictor terms included in the prior imputation model. Thus, an analysis model which regresses log-income against log working hours and the categorical predictors using a quadratic spline for the age main effect but only a linear age term in an age by sex interaction is congenial to the prior imputation model specified on the log-scale, whereas a similar analysis model including the non-linear components of the age spline term in the age by sex interaction would be uncongenial to the prior imputation model because the latter model included only the linear part of the age effect in two-way interactions. Similarly, an analysis model which regresses log income against untransformed working hours would be uncongenial to any of the prior imputation models considered because the only prior imputation models considered either modelled both weekly income and working hours on their original scale or modelled both variables on a log transformed scale.

When imputation and analysis models are congenial, multiple imputation theory predicts good inferential performance for multiply imputed data (Meng, 1994). Our use of hierarchical and non-parametric Bayesian imputation models is intended to improve robustness of inference using multiply imputed data to discrepancies between the prior imputation and analysis models and therefore the performance of these imputation methods when the prior imputation model structures are not congenial to the analysis models is the main focus of the investigation reported below. Accordingly, as can be seen from Table 3 we present results for several uncongenial models reflecting the different types of uncongeniality that can occur. These range from mild discrepancies such as the inclusion in the analysis model of additional components of age-related interactions not included in the prior imputation model, to more fundamental differences such as analysis models being specified on the log-scale while imputation models are specified on the original scale.

7.1.4 Model-fitting

The hierarchical Poisson log-linear imputation model for the categorical data was fitted using the PRIMM programme (Christiansen and Morris, 1997). The hierarchical Bayes imputation models for
the numerical variables, based on a multivariate normal regression as the centering model were fitted using the Gibbs Sampling algorithm outlined in 5.2.2. For each model we ran three independent chains and used the Gelman – Rubin convergence diagnostic to assess convergence (Gelman et al, 2004, p. 295-298). Burn-in periods of between 2000 and 4000 iterations were required in order to ensure Gelman-Rubin statistics less than 1.15 for all parameters. We saved 750 post-burn-in samples from the posterior distribution and then randomly sampled 100 of the stored parameter vectors for use in creating 100 synthetic datasets, as described in Section 6. All analysis models were fitted using the R glm() function (R Development Core Team, 2008).

7.1.5 Presentation of results

With several performance metrics to be reported, multiple subgroups and several analysis models on which to evaluate the synthetic data methods the volume of results is potentially large. We have therefore endeavoured to follow the advice of Gelman et al (2002) and report results graphically where possible. The figures reported below are typically arranged in a grid of graphs with rows of the grid corresponding to a particular performance metric and columns referring to a synthetic data method. Individual graphs within the grid contain a plot in which the value of a performance metric is plotted against a measure of precision and individual points represent either a subgroup (for descriptive statistics) or a model parameter. For descriptive statistics, the measure of precision used was the log (base 10) of the sample size. For model parameters the measure of precision was based on the standard error for the parameter estimate obtained from the real data (ie the CURF). However since standard errors varied considerably for different parameters we standardised the standard errors by dividing them by the corresponding point estimate obtained from the real data. This produced a measure of uncertainty which was generally comparable for different parameters, although anomalies can occur for parameters with small point estimates. We used a log-transformed version of the standardised standard error in the plots and in practice this provided a sensible index of uncertainty. Note that, because we used sample size as our index of precision for descriptive statistics and a standard error based measure of uncertainty for model parameter estimates, precision increases moving from left to right in the plots for descriptive statistics, but decreases moving left to right in the plots for model parameters. The measures of uncertainty and precision are primarily used as a device to spread out the values of the performance metrics over some range, but nevertheless do offer the opportunity to explore the possibility of trend in performance with increasing or decreasing precision.

For model parameters we use percentage absolute difference in point estimates defined as 100 \( \times \) (synthetic estimate – real estimate)/real estimate to quantify point estimate discrepancies rather than the absolute difference as used for descriptive statistics. This reflects the variation in magnitude of parameter estimates, reflecting, in part the different scales on which predictors are measured. For example the model parameter for the quadratic term for age is not comparable to the parameter for the binary sex indicator.

7.2 Descriptive statistics

We computed descriptive statistics for the overall CURF subset and for 35 subgroups defined by sex, ethnic group and educational achievement and sex by ethnic group and sex by educational achievement groups. The subgroups are those listed in Tables 1 and 2. These groups, which are not all mutually exclusive, served as the basis for the comparison of descriptive statistics. We compared the CURF descriptive statistics with the corresponding estimate obtained from multiply imputed synthetic versions of the CURF. Synthetic data point estimates of means were obtained by averaging over the estimates obtained from the 100 synthetic datasets and credible intervals were computed using the combined variance estimate and t-distribution approximation proposed by Raghunathan, Reiter and Rubin (2003). For medians and quartiles we used the median of the 100 synthetic estimates (medians or quartiles) as the synthetic data point estimate.

In Figures 1 and 2 we compare the performance of four synthetic data methods, in estimating average weekly income, using the CURF data as the real data. Moving from top to bottom the rows in the figures correspond to three different performance metrics: absolute difference in point estimates, relative credible interval length, comparing synthetic data intervals to the real data interval and the Karr et al data-utility measure based on credible interval overlap. In each case the performance metric is plotted on the vertical axis and the real data sample size is plotted on the horizontal scale, using a log (base 10) scale. Moving from left to right the four columns present results for the fully parametric hierarchical imputation model, a non-parametric hierarchical model with Dirichlet Process precision.
parameter \( \alpha = 19 \), a non-parametric hierarchical model with Dirichlet Process precision parameter \( \alpha = 9 \), and the Bayesian bootstrap method for numerical data.

Considering first the absolute differences in point estimates for imputations models fitted on the original scale (Figure 1, row 1), it can be seen that the scatter of differences exhibited by the fully parametric imputation model is reduced under the non-parametric approaches. Results for the non-parametric imputation models (labeled NPHB in the figures) with precision parameters \( \alpha = 9 \), \( \alpha = 19 \), were similar, with the major difference in absolute point estimate discrepancies being for the female, Māori combinations ethnic group (n= 165, shown as \( \log_{10}(165)=2.2 \) in the figures) for which the absolute discrepancy was $25.10 for the \( \alpha = 19 \) model, reducing to $12.00 for the \( \alpha = 9 \) model. Both these discrepancies were small relative to the standard error of the average income for this group in the CURF which was $33.90. Absolute differences between the Bayesian bootstrap and real data estimates were close to zero except for the two smallest groups where absolute differences of $41.80 and $33.10 were observed, respectively, for the male and female other combinations ethnic groups. The sample sizes for these groups in the CURF were n=22 for males and n=35 for females and the standard errors for average weekly income in these groups were $58.70 and $62.80 for males and females respectively. Thus the observed differences were somewhat smaller than one standard error. The discrepancy between the Bayesian bootstrap and CURF point estimates, is presumably due to Monte Carlo sampling error, illustrating that with small groups, even resampling directly from the data can lead to discrepancies in point estimates, depending on the frequency with which certain records are resampled. Small groups are also likely to be more affected than large groups by the need to supplement Bayesian bootstrap imputation with model-generated imputations for cells which are empty in the real data.

The rightmost point (largest sample size) in each of the plots corresponds to the overall population estimate. Absolute differences between synthetic data and CURF estimates were small relative to the average income of $711.10 for all four imputation models, but were largest for the fully parametric imputation model ($12.20) and smallest for the Bayesian bootstrap imputation method ($0.72). The absolute differences for the non-parametric hierarchical models were $2.81 for the \( \alpha = 19 \), and $2.33 for the \( \alpha = 9 \) case.

The relative lengths of credible intervals for the synthetic data estimates based on imputation modelling on the original scale (Figure 1, row 2) fluctuated either side of one with no clear relationship with sample size, except, possibly for the non-parametric model with \( \alpha = 9 \) for which there was some suggestion of increasing relative credible interval length with increasing sample size.

In terms of the credible interval overlap based measure of data utility, the final row in Figure one, shows some troubling performance for the fully parametric imputation model, particularly for the larger groups. This reflects the modest absolute differences in point estimates under the fully parametric imputation model, shown in the first row of Figure 1, being large relative to the standard errors for larger groups. For example the absolute difference of $12.18 between the synthetic data estimate obtained under the parametric imputation model and the CURF estimate is large relative to the CURF standard error of $4.36. In terms of posterior distributions this means that the shift in location is large relative to the spread resulting in a low value of 0.19 for the index of credibility overlap. To some extent this can be seen as an artifact of the large sample size of 11,267, since with large sample sizes standard errors are inevitably small and hence differences in measures of location have to be very small to ensure substantial overlap of posterior distributions. Nevertheless the other three imputation models compared in Figure 1 all achieved credible intervals close to the ideal value of 0.95, with the non-parametric model with \( \alpha = 9 \), in particular proving competitive against the Bayesian bootstrap method.

The performance of the imputation models for estimating mean income, when fitted to the CURF income and working hours data on the log-scale are compared in Figure 2. The pattern of performance is broadly similar to that shown in Table one with absolute point estimate differences being more consistently small and data-utility more consistently high for the non-parametric hierarchical imputation modelling approaches compared to the fully parametric imputation model. However, for both parametric and non-parametric modelling approaches, performance appears worse, for some groups, relative to the results obtained for imputation modelling on the original scale. This is not surprising since log scale imputation modelling can be expected to provide good estimates of the mean of log-transformed income, but since average income is not equal to the exponentiated value of average log income, good performance in estimation of the mean of log-transformed values does not necessarily translate to good performance in estimating the mean of the untransformed values. It is
therefore to be expected that the fully-parametric hierarchical model will be more adversely affected by moving to imputation modelling on the log-scale, than the non-parametric imputation models since imputations under this approach are fully model-dependent, whereas the non-parametric approaches enjoy the additional robustness due to the Polya sampling process.

Differences in the performance of the fully parametric and non-parametric hierarchical imputation models are most apparent for the credibility interval overlap based data utility results (row 3, Figure 2) where it can be seen that several low values of the data-utility index are recorded for the fully parametric imputation model. In particular there is a cluster of near-zero values in the bottom right-hand corner of the data utility plot for the parametric hierarchical model. These correspond to four large groups, the overall sample (n=11,267), all New Zealand Europeans (n=8665), all females (n=5,739) and female New Zealand Europeans (n=4420) with modest absolute point estimate differences which were nevertheless large relative to the standard errors obtained from the CURF. The absolute point estimate differences for these groups, with corresponding CURF standard errors in parentheses were $17.22 (4.36), $19.78 (5.12), $18.52 (4.70), $24.33 (5.44), respectively for the four groups listed above. Absolute point estimate differences for these groups under the non-parametric hierarchical imputation models and Bayesian bootstrap imputation method were negligible, and with credible interval lengths being comparable to those obtained from the CURF, data-utility was high for average income in these four large groups under the non-parametric hierarchical model and Bayesian bootstrap approaches.

The non-parametric hierarchical log scale imputation model with precision parameter $\alpha = 19$ produced an overestimate of average weekly income in the smallest group (Males, other combinations ethnic group, n= 22) of $52.11$, however all other point estimate differences were no greater than $22.03$. Absolute point estimate differences for average weekly income under the $\alpha = 9$ non-parametric log scale imputation model were no greater than $16.82$. Of more interest in the results for the non-hierarchical imputation models are two mid-sized groups with data-utility measures lying below the values for all other groups in the third row of Figure 2. These points correspond to the full Pacific ethnic group (n=541), and to the male subgroup of the Pacific ethnic group (n=290). Estimates of average weekly income for the latter group exceed the CURF estimate, by $22.03$ and $16.82$ for the non-parametric imputation models with $\alpha = 19$ and $\alpha = 9$, respectively. This over-estimation for the male Pacific group leads to similar differences for the full Pacific group (males and females combined) because estimates for the female Pacific group were close to the CURF estimate for both the non-parametric hierarchical imputation models. The point estimate differences were larger than the CURF standard errors of $13.57$ for the male Pacific group and $10.16$ for the overall Pacific group. The credibility interval overlap based utility measures were 0.66 ($\alpha = 19$) and 0.77 ($\alpha = 9$) for male, Pacific average weekly income and 0.69 ($\alpha = 19$) and 0.71 ($\alpha = 9$) for average weekly income in the overall Pacific group.

For all other groups, and, particularly for the $\alpha = 9$ case, data utility under the non-parametric hierarchical log scale imputation models approached the values achieved under Bayesian bootstrap imputation.
Comparison of MI methods: imputation modelling on original scale
Estimating mean income for 36 groups

Parametric HB  NPHB ($\alpha = 19$)  NPHB ($\alpha = 9$)  Bayes bootstrap

- **Absolute mean diff**
- **Relative CI length**
- **Data utility**

**Notes**
- Log scale for absolute mean difference
- Log scale for relative CI length
- Karr index scale

Comparison of estimating mean income for 36 groups using different MI methods.
Comparing MI methods: imputation modelling on log scale

Estimating mean income for 36 groups

<table>
<thead>
<tr>
<th>Method</th>
<th>Parametric HB</th>
<th>NPHB ($\alpha = 19$)</th>
<th>NPHB ($\alpha = 9$)</th>
<th>Bayes bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute mean diff</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
<td><img src="image4" alt="Graph" /></td>
</tr>
<tr>
<td>Relative CI length</td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
<td><img src="image7" alt="Graph" /></td>
<td><img src="image8" alt="Graph" /></td>
</tr>
<tr>
<td>Data utility</td>
<td><img src="image9" alt="Graph" /></td>
<td><img src="image10" alt="Graph" /></td>
<td><img src="image11" alt="Graph" /></td>
<td><img src="image12" alt="Graph" /></td>
</tr>
</tbody>
</table>

Differences between synthetic data and CURF point estimates of the median, lower quartile and upper quartile of the income distributions are shown in Figure 3, for imputation models fitted on the original scale and in Figure 4, for imputation models fitted on the log scale. Since the formulas for combining inferences across the synthetic datasets do not apply to medians and quartiles, only point estimate differences, treating the CURF estimate as the gold standard, are compared in Figures 3 and 4.

Performance of the synthetic data methods was similar for imputation models fitted on the original and on the log-scale. The most striking features of Figures 3 and 4 are the smaller point estimate differences at large sample sizes for the non-parametric imputation models compared to fully parametric imputation. This reflects the lower weight given to the hierarchical multivariate normal model as cell size increases, under the non-parametric imputation models. Despite the superior performance of the non-parametric approaches at larger sample sizes some important discrepancies remain for smaller groups. For example, the differences between synthetic data and CURF estimates for the upper income quartile for the female, other combinations ethnic group ($n=35$) were $97 and $105 per week under the models with $\alpha = 19$ and $\alpha = 9$, respectively.
Figure 3

Comparison of MI methods: imputation modelling on original scale

Estimating income quartiles for 36 groups

<table>
<thead>
<tr>
<th>Method</th>
<th>Plot 1</th>
<th>Plot 2</th>
<th>Plot 3</th>
<th>Plot 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric HB</td>
<td><img src="image1.png" alt="Plot" /></td>
<td><img src="image2.png" alt="Plot" /></td>
<td><img src="image3.png" alt="Plot" /></td>
<td><img src="image4.png" alt="Plot" /></td>
</tr>
<tr>
<td>NPHB ($\alpha = 19$)</td>
<td><img src="image5.png" alt="Plot" /></td>
<td><img src="image6.png" alt="Plot" /></td>
<td><img src="image7.png" alt="Plot" /></td>
<td><img src="image8.png" alt="Plot" /></td>
</tr>
<tr>
<td>NPHB ($\alpha = 9$)</td>
<td><img src="image9.png" alt="Plot" /></td>
<td><img src="image10.png" alt="Plot" /></td>
<td><img src="image11.png" alt="Plot" /></td>
<td><img src="image12.png" alt="Plot" /></td>
</tr>
<tr>
<td>Bayes bootstrap</td>
<td><img src="image13.png" alt="Plot" /></td>
<td><img src="image14.png" alt="Plot" /></td>
<td><img src="image15.png" alt="Plot" /></td>
<td><img src="image16.png" alt="Plot" /></td>
</tr>
</tbody>
</table>

Comparison of MI methods: imputation modelling on original scale

Estimating income quartiles for 36 groups

<table>
<thead>
<tr>
<th>Method</th>
<th>Plot 1</th>
<th>Plot 2</th>
<th>Plot 3</th>
<th>Plot 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric HB</td>
<td><img src="image1.png" alt="Plot" /></td>
<td><img src="image2.png" alt="Plot" /></td>
<td><img src="image3.png" alt="Plot" /></td>
<td><img src="image4.png" alt="Plot" /></td>
</tr>
<tr>
<td>NPHB ($\alpha = 19$)</td>
<td><img src="image5.png" alt="Plot" /></td>
<td><img src="image6.png" alt="Plot" /></td>
<td><img src="image7.png" alt="Plot" /></td>
<td><img src="image8.png" alt="Plot" /></td>
</tr>
<tr>
<td>NPHB ($\alpha = 9$)</td>
<td><img src="image9.png" alt="Plot" /></td>
<td><img src="image10.png" alt="Plot" /></td>
<td><img src="image11.png" alt="Plot" /></td>
<td><img src="image12.png" alt="Plot" /></td>
</tr>
<tr>
<td>Bayes bootstrap</td>
<td><img src="image13.png" alt="Plot" /></td>
<td><img src="image14.png" alt="Plot" /></td>
<td><img src="image15.png" alt="Plot" /></td>
<td><img src="image16.png" alt="Plot" /></td>
</tr>
</tbody>
</table>
Figure 4

Comparison of MI methods: imputation modelling on log scale

Estimating income quartiles for 36 groups

<table>
<thead>
<tr>
<th>parametric HB</th>
<th>NPHB ($\alpha = 19$)</th>
<th>NPHB ($\alpha = 9$)</th>
<th>Bayes bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference in medians</td>
<td>Difference in medians</td>
<td>Difference in medians</td>
<td>Difference in medians</td>
</tr>
<tr>
<td>$\log_{10}(n)$</td>
<td>$\log_{10}(n)$</td>
<td>$\log_{10}(n)$</td>
<td>$\log_{10}(n)$</td>
</tr>
<tr>
<td>0 40 80</td>
<td>0 40 80</td>
<td>0 40 80</td>
<td>0 40 80</td>
</tr>
<tr>
<td>1.5 2.5 3.5</td>
<td>1.5 2.5 3.5</td>
<td>1.5 2.5 3.5</td>
<td>1.5 2.5 3.5</td>
</tr>
<tr>
<td>Difference in LQ</td>
<td>Difference in LQ</td>
<td>Difference in LQ</td>
<td>Difference in LQ</td>
</tr>
<tr>
<td>$\log_{10}(n)$</td>
<td>$\log_{10}(n)$</td>
<td>$\log_{10}(n)$</td>
<td>$\log_{10}(n)$</td>
</tr>
<tr>
<td>0 40 80</td>
<td>0 40 80</td>
<td>0 40 80</td>
<td>0 40 80</td>
</tr>
<tr>
<td>1.5 2.5 3.5</td>
<td>1.5 2.5 3.5</td>
<td>1.5 2.5 3.5</td>
<td>1.5 2.5 3.5</td>
</tr>
<tr>
<td>Difference in UQ</td>
<td>Difference in UQ</td>
<td>Difference in UQ</td>
<td>Difference in UQ</td>
</tr>
<tr>
<td>$\log_{10}(n)$</td>
<td>$\log_{10}(n)$</td>
<td>$\log_{10}(n)$</td>
<td>$\log_{10}(n)$</td>
</tr>
<tr>
<td>0 40 80</td>
<td>0 40 80</td>
<td>0 40 80</td>
<td>0 40 80</td>
</tr>
<tr>
<td>1.5 2.5 3.5</td>
<td>1.5 2.5 3.5</td>
<td>1.5 2.5 3.5</td>
<td>1.5 2.5 3.5</td>
</tr>
</tbody>
</table>

LQ = lower quartile; UQ = upper quartile

7.3 Log-linear regression analysis models

7.3.1 Congenial models

Figures 5 and 6 summarise the performance of the synthetic data methods with respect to two analysis models which are congenial to the prior imputation model specified on the log-scale. The analysis models are log-linear regressions of income on log-working hours and categorical predictors, with Figure 5 summarising results for model including only main effect terms and Figure 6 summarising results for a model including main effects plus a single interaction term defined as the linear part of the age effect multiplied by the binary sex indicator. Quadratic splines with a single knot at age 45 are used to represent age main effects in both analysis models, as is also the case for the prior imputation model.

Each rows of plots in the figures provide information on a particular performance metric, by plotting, for each model parameter, the value of performance metric against a measure of precision. As discussed in 7.1.5, we used standardised standard errors, obtained by dividing the real data standard
errors by the real data point estimate, as our index of precision and plot this on the log (base 10) scale. A value of zero on this scale indicates a parameter for which the real data standard error is equal to the real data point estimate. Hence values greater than zero on this scale indicate parameters which are, relatively, imprecisely estimated in the real data and values substantially below zero indicate parameters which are relatively precisely estimated in the real data.

The first row of plots in the Figures 5 and 6, and all subsequent similar figures, summarises the discrepancy in point estimates between the synthetic data and real data (CURF, in this case). As a measure of this discrepancy we used the absolute percentage difference defined as $100 \times \frac{\text{abs(synthetic estimate – real estimate)}}{\text{real estimate}}$. The second row in the figures summarises the relative credible interval length for synthetic credible intervals compared to the real data intervals, for each model parameter. The final row in each plot presents Karr’s data utility measure based on credible interval overlap.

For both congenial analysis models performance is similar across the four synthetic data methods. Point estimate discrepancies with the real data estimate are generally small but are large for the least precisely estimated parameter, the Māori combinations ethnic group indicator (the rightmost point in the plots). However because the absolute differences in point estimates for this parameter are of the order of one real data standard error (0.02 to 0.05) and credible interval lengths are comparable for synthetic and real data the point-estimate discrepancies have negligible impact on credible interval overlap (data utility). The credible interval based data utilities for this parameter in the main effects model ranged from 0.891 (non-parametric Bayes hierarchical imputation model, $\alpha = 19$ ) to 0.942 (Bayesian bootstrap imputation) and from 0.895 (non-parametric Bayes hierarchical imputation model, $\alpha = 19$ ) to 0.943 (Bayesian bootstrap imputation) for the model including the linear age by sex interaction (Figure 6). Data utilities were high for all other parameters in the two congenial models. However, under the non-parametric hierarchical and Bayesian bootstrap imputation models, the data utilities for the two (relatively) most precisely estimated parameters, the intercept and log-working hours parameters, were approximately 0.8 for both models, due to some inflation of credible interval lengths, which is clearly apparent in the second row of Figure 5 and Figure 6.
Figure 5

Comparison of MI methods: Imputation modelling on log scale.
Estimating income main effects log-linear regression, working hours log transformed.

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter</th>
<th>Absolute % difference</th>
<th>Relative CI length</th>
<th>Data utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric HB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPHB ($\alpha = 19$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPHB ($\alpha = 9$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayes bootstrap</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
7.3.2 Models that are uncongenial due only to the inclusion of non-linear component of the age-sex interaction

Figure 7 summarises the performance of the log-scale imputation models for estimating a log-linear regression for income on log hours worked, categorical predictors and two-way interactions as described in Table 3. This analysis model includes the full quadratic spline representation of age in an age by sex interaction, as well as the main effect for age. In practice this means that the age-sex interaction involves two terms which are not included in the prior imputation model, which included only the linear part of the age effect in the age by sex interaction. The non-linear components of the quadratic spline for age are a quadratic age term and an additional term defined as zero for ages less than 45 years and as the squared difference between age and 45 years, for ages greater than or equal to 45. Each of these non-linear terms, as well as the linear age effect was multiplied by the binary sex indicator to produce the age by sex interaction in the analysis model, for which results are summarised in Figure 7.

In this mildly uncongenial scenario the fully-parametric hierarchical model achieved high data utilities for a high proportion of the 48 model parameters and the non-parametric hierarchical imputation models further improved on this good performance. For example the proportion of the 48 model parameters for which data utility of at least 0.9 was achieved was 62.5%, 66.7%, 68.8% and 66.7%,
for the fully parametric, non-parametric ($\alpha = 19$), non-parametric ($\alpha = 9$) and Bayesian bootstrap imputation models, respectively. The corresponding proportions of parameters for which data utilities of at least 0.8 were achieved were 83.3%, 89.6%, 87.5%, 81.2% and the minimum data utilities were 0.395, 0.474, 0.598, 0.597. Thus it appears that the performance of the four imputation methods was similar with respect to the number of parameters for which reasonable credible interval overlap was achieved but the non-parametric hierarchical imputation model with $\alpha = 9$ and Bayesian bootstrap imputation provided better protection against low credible interval overlap. The lowest credible interval overlap indices (data utility) occurred, not for parameters related to the age-sex interaction, but to ethnic group parameters and the corresponding low data utilities appear to due principally due to moderate point estimate discrepancies. Substantial credible interval length inflation for some parameters was apparent under all four imputation models (second row, Figure 7). The parameters most affected by this were the intercept, the log-working hours main effect parameter, and some interaction terms involving this parameter, the sex and ethnic group main effect parameters, particularly the indicators for the two smallest ethnic groups, Māori combinations and other combinations.
7.3.3 Models that are uncongenial due only to differences in the scale on which models are specified

The multivariate normal prior imputation model implies a linear relationship between components on the scale on which the model is specified. Thus a prior multivariate normal model for income and working hours specified on the original scale implies a linear relationship between weekly income and working hours, whereas a prior multivariate normal specified on the log-scale implies a linear relationship between log weekly income and log working hours.

For the log-linear regression models for income and the prior imputation models considered in this report uncongeniality due to differences in the scale on which the analysis and prior imputation models are specified can arise in three different ways. Firstly the scale for the outcome variable (weekly income) may be differ between analysis and prior imputation models, secondly the scale used for the predictor variable, hours worked may differ and thirdly, the scale for both weekly income and working hours may differ. The performance of the synthetic data methods for each of these scenarios is summarised in Figures 8 to 10.
Figure 8 illustrates the performance of the synthetic data methods when only the scale on which the outcome variable is modelled differs between the analysis model and prior imputation model. The imputation models underlying Figure 8 were specified on the original scale for weekly income (dollars) and hours worked (hours) whereas the analysis model was a main effects regression of log-income on working hours and the categorical predictors, age-group, sex, ethnicity and qualifications. The main difference in the performance of the imputation methods for this scenario concerns four parameters clearly shown as having low data utility under the fully-parametric hierarchical Bayes imputation model. In order from lowest to highest data utility (under the fully parametric imputation model) the parameters are working hours (data utility, 0.000), Pacific ethnicity (0.315), intercept (0.354) and university qualifications (0.608). The data utilities show a clear improvement under the non-parametric hierarchical Bayesian and Bayesian bootstrap models, with only the working hours and Pacific ethnicity parameters remaining below 0.8 under the non-parametric hierarchical Bayes imputation models. For the non-parametric imputation model with Dirichlet Process precision parameter $\alpha = 19$, the data utilities for working hours and Pacific ethnicity parameters were 0.553 and 0.706, respectively and the corresponding data utilities under the non-parametric model with $\alpha = 9$ were 0.745 and 0.700. Under the Bayesian bootstrap imputation data utilities for the working hours and Pacific parameters improved further to 0.824 and 0.917.
Comparison of MI methods: imputation modelling on original scale

Estimating income main effects log-linear regression, work hours untransformed

parametric HB      NPHB ($\alpha = 19$)      NPHB ($\alpha = 9$)      Bayes bootstrap

**Figure 8**
Figure 9 illustrates the performance of the synthetic data methods under a scenario where uncongeniality between analysis and prior imputation models arises solely from a difference between the scale on which a predictor variable is represented in the analysis prior imputation models. The imputation model underpinning Figure 9 was specified on the log-scale for both weekly income and hours worked, whereas the analysis model for which results are reported was a log-linear regression of income against actual working hours and the categorical data predictors. Thus, in this scenario, the log-linear relationship between weekly income and hours worked assumed by the analysis model is mis-modelled by the prior imputation model as linear relationship between log weekly income and log-working hours.

From Figure 9 it can be clearly seen that data utility was extremely low for three parameters under both fully parametric and non-parametric hierarchical imputation models. The three parameters in question are the intercept, the working hours parameter and the sex parameter. The non-parametric hierarchical imputation models appear not to have provided any additional robustness to this type of mis-modelling of the income-working hours relationship, relative to the fully parametric imputation model. These low data utilities result primarily from point estimate discrepancies which although not large in absolute terms are large relative to the real data standard errors. Taking the fully parametric hierarchical imputation model as an example, the point estimate discrepancies for the intercept, working hours parameter and sex parameter were equivalent to 10.6, 35.7 and 13.3 real data standard errors, respectively. In addition, in the case of the working hours parameter there was substantial inflation of credible interval length, under the fully parametric and non-parametric hierarchical imputation models and somewhat less so under the Bayesian bootstrap imputation. Overall, the results indicate the hierarchical imputation models based on a prior model which posits a linear relationship between log weekly income and log hours worked are not robust against analysis models which specify a linear relationship between log weekly income and hours worked.

The poor performance with respect to the intercept and sex parameters, in addition to the working hours parameter is likely to be due to correlations between these parameters. When working hours was removed from the analysis model good performance for estimating the remaining parameters was observed as summarised in Figure 10. This finding is unsurprising because with working hours removed from the model the analysis model is congenial to the log-scale prior imputation model.

Figure 11 summarises results for a third scale related uncongenial scenario in which a log-linear regression of income against log working hours and categorical predictors is estimated from synthetic data based on the prior imputation model which uses both weekly income and hours worked per week on their original scale. In this scenario the non-parametric hierarchical imputation models improve on the fully parametric imputation model. Data utilities close to zero were recorded for the intercept and working hours parameters under the full parametric hierarchical imputation model and data utility for these parameters improved to at least 0.6 under the non-parametric imputation models.
Comparison of MI methods: Imputation modelling on log scale.

Estimating income main effects log-linear regression, working hours untransformed.

Parametric HB  NPHB ($\alpha = 19$)  NPHB ($\alpha = 9$)  Bayes bootstrap

Absolute %difference.

Relative ci length

Data utility

log standardised std. error

Karr index

-2.5 -1.0 0.5

0.0 0.4 0.8

% difference

% difference

% difference

% difference

-2.5 -1.0 0.5

0 50 150

-2.5 -1.0 0.5

0 50 150

-2.5 -1.0 0.5

0 50 150

-2.5 -1.0 0.5

0.0 0.4 0.8

-2.5 -1.0 0.5
Figure 10

Comparison of MI methods: Imputation modelling on log scale.
Estimating income main effects, log-linear regression working hours omitted

Parametric HB  NPHB ($\alpha = 19$)  NPHB ($\alpha = 9$)  Bayes bootstrap

- Absolute % difference
- Relative ci length
- Data utility

log standardised std. error  log standardised std. error  log standardised std. error  log standardised std. error
Figure 11

Comparison of MI methods: imputation modelling on original scale

Estimating income main effects log-linear regression, work hours log transformed

parametric HB  NPHB (α = 19)  NPHB (α = 9)  Bayes bootstrap

Absolute pct. difference

Relative ci length

Data utility

Confidentialising Microdata

Official Statistics Research Series, 5, 2009
www.statisphere.govt.nz
7.3.4 Models that are uncongenial due to differences in the scale used for model specification and inclusion of non-linear components of the age-sex interaction term

Prior imputation model specified on original scale; analysis model regresses log-income on untransformed working hours, categorical predictors and interactions including non-linear components of the age by sex interaction.

Figure 12 summarises the performance of imputation models, specified on the original scale for both weekly income and hours worked, in estimating a log-linear regression model for income against working hours (on the original scale) and categorical predictors, including two-way interactions as described for model 9 in Table 3. The analysis model departs from the prior imputation model in two respects. Firstly the analysis model models income on the log-scale rather than on the original scale as in the prior imputation model and secondly the analysis model includes among its predictors the non-linear components of the quadratic spline for age in and age by sex interaction whereas the prior imputation model includes only the linear part of the age effect in and age by sex interaction.

Figure 12 clearly shows non-parametric hierarchical imputation models exhibited better performance than the fully parametric imputation model, for this type of uncongeniality. Reading the Figure from left to right, improvements in all three performance metrics can be seen, and the performance of the non-parametric hierarchical model with $\alpha = 9$ is similar to that achieved under Bayesian bootstrap imputation. The proportion of the 48 model parameters for which the synthetic data methods achieved data utilities of at least 0.8 were, 0.48, 0.69, 0.79 and 0.83, respectively for the fully-parametric hierarchical imputation model, the non-parametric hierarchical models with $\alpha = 19$, and $\alpha = 9$, and the Bayesian bootstrap imputation. The minimum data-utility achieved by the four synthetic data methods was 0.00 (fully parametric HB), 0.60 (non-parametric HB, $\alpha = 19$), 0.67 ($\alpha = 9$), and 0.75 (Bayesian bootstrap). The data-utility results indicate improving performance as imputation models move from fully parametric to the fully non-parametric Bayesian bootstrap. The non-parametric hierarchical imputation models and Bayesian bootstrap therefore avoided the very low data utilities sometimes observed for the fully parametric hierarchical imputation model.

Substantial point estimate discrepancies were observed for the ‘Other’ ethnic group by secondary qualifications term which is shown as the right most point in the plots. In addition, under the fully parametric hierarchical imputation model, but not the other imputation models a large point estimate discrepancy was also observed for the other combinations ethnic group by trade qualifications interaction term, shown as the second point from the right. However, because these parameters are poorly estimated, in terms of precision, the large point estimate discrepancy has little impact on credible interval overlap.
Prior imputation model included only the linear part of the age effect in an age by sex interaction.

Turning now to the opposite form of congeniality to that reported on immediately above, Figure 13 summarises the performance of imputation models, under an uncongenial scenario in which the prior imputation model is specified on the log scale for both income and hours worked, the analysis model also models income on the log-scale but disagrees with the prior imputation model by using untransformed hours worked as a predictor. In addition the analysis model reported on in Figure 13 includes non-linear components of the age quadratic spline in the age by sex interaction, whereas the prior imputation model included only the linear part of the age effect in an age by sex interaction.

Overall the performance of the imputation methods for this scenario is reasonable for the majority of parameters, and there appear to be clear improvements in data utility under the non-parametric hierarchical imputation models compared to the fully parametric hierarchical imputation model. For example the proportion of the 48 parameters for which the credible interval overlap based data utility
Confidentialising Microdata

index exceeded 0.8 was 0.62 for the fully parametric imputation model, 0.79 under the non-parametric model with $\alpha = 19$, 0.83 under the non-parametric model with $\alpha = 9$ and 0.92 for the Bayesian bootstrap. The corresponding proportions for data utility exceeding 0.9 were 0.44, 0.60, 0.71 and 0.77, respectively for the four imputation methods. However, a concern with the fully parametric hierarchical imputation model is that data utilities were less than 0.2 for seven parameters and between 0.2 and 0.4 for a further three parameters. These low data utilities occurred for the intercept parameter, the sex and working hours parameters and for some interaction terms involving sex and working hours. It can also be seen from Figure 13 that the low data utilities occurred for parameters which are, relatively, precisely estimated (located to the left of the plots) and appear primarily due to small to moderate point estimate discrepancies which are large relative to the standard errors involved. The situation regarding parameters with very low data utilities improves for the non-parametric hierarchical imputation models with only three parameters (intercept, working hours and sex main effects) having data utilities less than 0.2 under non-parametric model with $\alpha = 19$, and only one parameter (working hours main effect) having data utility less than 0.2, under the non-parametric imputation model with $\alpha = 9$. 


Prior imputation model specified on original scale; analysis model specified as regression of log income on log working hour, categorical predictors and interactions, including non-linear age terms in an age by sex interaction.

In the final uncongenial scenario considered, both weekly income and working hours were mis-modelled by the prior imputation model relative to the analysis model, and in addition, the analysis model included non-linear age components of the age effect in the age by sex interaction, whereas the prior imputation model included only the linear part of the age quadratic spline in the age by sex interaction. In this scenario the prior imputation model was specified on the original scale, implying a linear relationship between weekly income and working hours, while the analysis model regressed log weekly income against log working hours, categorical predictors and interaction terms as described for model 6, in Table 3. The performance of the synthetic data methods under this scenario is summarised in Figure 14.

While some large percentage point estimate discrepancies were observed, these were for poorly estimated parameters for which real data estimates were close to zero. The large point estimate discrepancies had little impact on credible interval overlap with data utilities close to 0.95 under all imputation models for the parameters plotted as the rightmost points (biggest standardised standard errors) in Figure 14. The non-parametric hierarchical imputation models clearly outperformed the fully parametric imputation models with respect to data utility. While the minimum data utilities under the
former models were 0.60 and 0.67, for the $\alpha = 19$ and $\alpha = 9$ models respectively, data utility was less than 0.4 for seven parameters and close to zero (less than 0.001) for two parameters (intercept and working hours main effect) under the fully parametric hierarchical imputation model.

Figure 14

Comparison of MI methods: imputation modelling on original scale
Estimating income log-linear regression with interactions, work hours log transformed
parametric HB  NPHB ($\alpha = 19$)  NPHB ($\alpha = 9$)  Bayes bootstrap
7.4 Disclosure risks

Assessment of disclosure risks for multiply imputed synthetic data is an under-developed area although approaches for fully synthetic data and partially synthetic data in which only selected variables are imputed are beginning to appear in the literature (Reiter, 2005; Dreschler et al, 2008; Reiter & Mitra, 2009). Here we take a simple approach and concentrate on predictive disclosure risk for weekly income. Specifically we investigate whether the multiply imputed synthetic data can be straightforwardly used to estimate the income of individuals known (by an intruder) to be in the survey dataset and to have certain characteristics which are accurately recorded in the survey data. This is a worst-case scenario since it requires an intruder to have detailed knowledge about particular individuals before interrogating confidentialised data.

Official statistics agencies are often particularly concerned about disclosure risks for so-called sample uniques. These are the individuals who are uniquely defined, in the real dataset, by a given set of categorical variables. Consequently, in order to investigate predictive disclosure risks associated with multiply imputed synthetic data, we concentrate here on predictive disclosures for the 54 individuals in the CURF dataset who are uniquely defined by age group (in five year age categories), sex, ethnic group and educational achievement level. The CURF dataset is again treated as the real dataset for this exercise and we use multiply imputed synthetic versions of the CURF to estimate weekly income levels of individuals as recorded in the CURF. We concentrate on the scenario in which an intruder knows the age, sex, ethnic group and educational achievement level of one or more survey respondents and endeavours to estimate their income by using the cell medians computed from the multiply imputed synthetic datasets. Note that for cells with a real count of one, synthetic cell counts may be zero, one or greater than one.

Figures 15 and 16 show the distribution of the median prediction errors, defined as the difference between the median of 100 synthetic data medians and the real data median, by log (base 10) cell size. Median prediction errors for imputation models specified on the original scale are shown in Figure 15 and the corresponding results for imputation models specified on the log-scale are shown in Figure 16. The leftmost strip of data in the plots gives the distribution for the 54 cells with a count of one. In this case the real data median is just the recorded income for the single individual included in the cell and the median error is the difference between the median of the synthetic data median incomes for the cell and the individual’s actual recorded income. The distribution of median errors appears similar under the original scale and log-scale prior imputation model and the errors in the n=1 case spans a wide range. Note that the errors converge to zero as the sample size increases. This is desirable as accurate estimation for large cells is important order to ensure good inferential performance for descriptive statistics and model parameters. For large cell sizes accurate estimation of cell medians would not usually pose disclosure risks because of variation of individual responses within a cell.
Figure 15

Comparison of median prediction errors for 405 cells, by cell size
Imputation models fitted on original scale

fully parametric

non-parametric, ($\alpha = 19$)

non-parametric, ($\alpha = 9$)
The distribution of median prediction errors for the 54 unique individuals is explored in more detail in, Tables 4 and 5, where, in Table 4, the distribution of the errors is described for each of the synthetic data methods and, in Table 5, the proportion of the 54 individuals for whom synthetic data estimates fall within $10, $20 or $50 of the actual value are presented. All synthetic data methods, except the Bayesian bootstrap appear to give a high degree of protection for income values, in the sense that an intruder would have to tolerate a large amount of error if using synthetic data estimates to infer actual income for one of the 54 unique values.

It can be seen from Tables 4 and 5 that Bayesian bootstrap imputation exactly reproduces the responses of the unique individuals. This reflects the fact that the Bayesian bootstrap approach resamples from the observed data and for cells containing a single individual that individual’s data is therefore drawn for each imputation. This illustrates why Bayesian bootstrap imputation is not a realistic option as a general imputation method for synthetic data.

The parametric and non-parametric hierarchical Bayes imputation models produced similar distributions of median prediction errors for the 54 unique individuals. This may seem surprising given
that the fully parametric model corresponds to a Dirichlet Process precision parameter of $\alpha = \infty$. However, because the differences described in the Table 4 refer only to cells with a real count of one, the cell median estimates obtained from the non-parametric hierarchical Bayes models rely heavily on the parametric centering model, since for cells with an observed cell count of one the probability of drawing the $i^{th}$ synthetic value from the fitted centering model is $\pi_i = \alpha / (\alpha + i)$. Moreover, since synthetic cell counts for cells with a true count of one will tend to be small the $\pi_i$ will be close to one for each synthetic value drawn because $i$ will be small. For example, for $\alpha = 9$, the probabilities of drawing from the model when imputing synthetic incomes for a cell with a true count of one and synthetic count of three are 0.9, 0.82, and 0.75, respectively for the three draws required. Recalling that under Polya urn sampling if a value is drawn from the model it is added to the data-urn for future draws, the likelihood of drawing model generated values, rather than the single real data point, on the second and third draws in the above example is greater than indicated by the stated probabilities.

Table 4

<table>
<thead>
<tr>
<th>Imputation Model</th>
<th>Min</th>
<th>25th percentile</th>
<th>Median</th>
<th>75th percentile</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric HB$^1$</td>
<td>-582.10</td>
<td>-172.20</td>
<td>-43.26</td>
<td>132.23</td>
<td>3034.00</td>
</tr>
<tr>
<td>Parametric HB, log$^2$</td>
<td>-524.70</td>
<td>-152.80</td>
<td>11.57</td>
<td>116.0</td>
<td>3286.00</td>
</tr>
<tr>
<td>NPHB, $\alpha = 19$ $^3$</td>
<td>-569.00</td>
<td>-141.70</td>
<td>-4.91</td>
<td>131.80</td>
<td>2385.00</td>
</tr>
<tr>
<td>NPHB, $\alpha = 19$, log$^4$</td>
<td>-464.80</td>
<td>-118.70</td>
<td>8.28</td>
<td>126.30</td>
<td>3232.00</td>
</tr>
<tr>
<td>NPHB, $\alpha = 9$ $^5$</td>
<td>-541.30</td>
<td>-160.80</td>
<td>0.01</td>
<td>145.20</td>
<td>2897.00</td>
</tr>
<tr>
<td>NPHB, $\alpha = 9$, log$^6$</td>
<td>-471.80</td>
<td>-123.70</td>
<td>-8.89</td>
<td>106.40</td>
<td>3213.00</td>
</tr>
<tr>
<td>Bayes bootstrap $^7$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

1 Fully parametric hierarchical Bayesian imputation model for numerical data fitted on the original scale.
2 Fully parametric hierarchical Bayes imputation model for numerical data fitted on the log scale.
3 Non-parametric hierarchical Bayes imputation model with Dirichlet Process precision parameter $\alpha = 19$, fitted on the original scale.
4 Non-parametric hierarchical Bayes imputation model with Dirichlet Process precision parameter $\alpha = 19$, fitted on the log scale.
5 Non-parametric hierarchical Bayes imputation model with Dirichlet Process precision parameter $\alpha = 9$, fitted on the original scale.
6 Non-parametric hierarchical Bayes imputation model with Dirichlet Process precision parameter $\alpha = 9$, fitted on the log scale.
7 Bayesian bootstrap imputation.

Table 5

<table>
<thead>
<tr>
<th>Imputation Model</th>
<th>Proportion within $$10$</th>
<th>Proportion within $$20$</th>
<th>Proportion within $$50$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric HB$^1$</td>
<td>0.0%</td>
<td>5.6%</td>
<td>20.4%</td>
</tr>
<tr>
<td>Parametric HB, log$^2$</td>
<td>3.7%</td>
<td>3.7%</td>
<td>18.5%</td>
</tr>
<tr>
<td>NPHB, $\alpha = 19$ $^3$</td>
<td>7.4%</td>
<td>7.4%</td>
<td>20.4%</td>
</tr>
<tr>
<td>NPHB, $\alpha = 19$, log$^4$</td>
<td>5.6%</td>
<td>11.1%</td>
<td>20.4%</td>
</tr>
<tr>
<td>NPHB, $\alpha = 9$ $^5$</td>
<td>13.0%</td>
<td>18.5%</td>
<td>20.4%</td>
</tr>
<tr>
<td>NPHB, $\alpha = 9$, log$^6$</td>
<td>3.7%</td>
<td>13.0%</td>
<td>24.1%</td>
</tr>
<tr>
<td>Bayes bootstrap $^7$</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

1 Fully parametric hierarchical Bayesian imputation model for numerical data fitted on the original scale.
2 Fully parametric hierarchical Bayes imputation model for numerical data fitted on the log scale.
3 Non-parametric hierarchical Bayes imputation model with Dirichlet Process precision parameter $\alpha = 19$, fitted on the original scale.
4 Non-parametric hierarchical Bayes imputation model with Dirichlet Process precision parameter $\alpha = 19$, fitted on the log scale.
5 Non-parametric hierarchical Bayes imputation model with Dirichlet Process precision parameter $\alpha = 9$, fitted on the original scale.
6 Non-parametric hierarchical Bayes imputation model with Dirichlet Process precision parameter $\alpha = 9$, fitted on the log scale.
7 Bayesian bootstrap imputation.
8 Comparison of multiply imputed synthetic data, sufficiency based perturbation and current Statistics New Zealand confidentialising methods

8.1 Design of the comparison

In this section we compare the performance of multiply imputed synthetic data, using the non-parametric hierarchical Bayesian imputation model described in section 4.5, with the performance of the CURF and another recently proposed confidentialising method known as sufficiency based perturbation (SBP) or simply perturbation (Muralidhar and Sarathy, 2006). For this exercise, the original IS2003 data is used as the gold standard and inferences obtained from the CURF, synthetic and perturbed versions of the IS2003 are compared with inferences obtained directly from the survey data. The comparisons are based on the same set of descriptive statistics and models as used in the example discussed in section 7, which used the CURF as the gold-standard. The same restrictions applied to define the CURF subset used for the comparisons reported in section 7 were applied to the original IS2003 data. That is, the dataset was restricted to people with positive income, who worked at least one-hour per week and were aged between 25 and 64.

The sufficiency based perturbation method (SBP) of Muralidhar and Sarathy (2006) deals only with numerical variables and seeks to preserve sufficient statistics for standard statistical analyses which assume (multivariate) normality. In particular the method preserves sufficient statistics for a multivariate linear regression of numerical variables on categorical variables through clever perturbation of residuals. The method produces confidentialised data which exactly preserves the multivariate linear regression fitted to the original data. That is, the parameter estimates, standard errors and covariances for the parameter estimates obtained from fitting the multivariate linear regression model to the perturbed data are identical to those obtained from fitting the model to the original data. As a consequence, the overall sample mean of modelled numerical variables is identical in the original and perturbed data as are sub-group means, for each sub-group, up to the level of detail indicated by the predictors in the model. Sample means are not preserved for from stratifications which are finer than those represented in the model. For example, perturbed data derived from a model incorporating two way-interactions between categorical predictors, but no higher order interactions will not preserve sample means for sub-groups formed by three way stratification of the predictors, but will preserve means for all groups defined by two-way stratification. Being based on a multivariate linear regression model, the perturbation method preserves the residual variance matrix, under the homoscedasticity assumption that the variance is constant over values of $X$. That is, for a given multivariate linear regression model the sample estimate of $V(Y|X)$ is identical for the original and perturbed data, for a given multivariate linear regression model, under the assumption that $V(Y|X)$ is constant in $X$. This leads to preservation of the overall sample variance but does not usually preserve sub-group variances for all subgroups because, in practice, the homoscedasticity assumption does not hold exactly.

As discussed in section 7, Statistics New Zealand uses a variety of confidentialising techniques to construct the CURF. For numerical variables the principle modification is top-coding, and sometimes bottom-coding, of extreme values, which involves replacing extreme values, such as those lying beyond the 99.5th percentile with the average of the extreme values. This strategy preserves sample means. Top and bottom coding is carried out with respect to the overall distribution rather than within strata and hence may not preserve sub-group means. As an additional measure to minimise disclosure risk, some records deemed to be of high disclosure risk, are deleted from the CURF.

The prior model structure used in the hierarchical centering model for the non-parametric hierarchical Bayes imputation model and for the model underpinning the perturbed data (henceforth the perturbation model) was identical to the structure specified as the prior imputation model in section 7. That is, the predictors in prior imputation and perturbation models included main effect terms for age-group, sex, ethnicity and educational achievement level and two-way interactions between these variables. A quadratic spline with a single knot at age 45 was used to model the age effects with each age-group represented by the corresponding category midpoint. Only the linear part of the age-effect was included in age-related interaction terms. All variables other than age were represented using dummy-variable coding in which one category is treated as a reference category and other categories represented as indicator variables. Imputation and perturbation models were fitted to the weekly income and working hours data on both their original scale and to (natural) log-transformed versions of these variables. We refer to the former as the ‘original scale’ models and the latter as ‘log-scale’ models.
For the multiply-imputed synthetic data a hierarchical Bayes log-linear Poisson model was used as the imputation model for categorical data. The prior model specification for this model was identical to that described in section 7 for modelling the CURF data and included all two-way interactions between the categorical variables. For the perturbed data, cell counts for cross-classified variables were equal to the observed counts since the perturbation method only modifies numerical variables.

### 8.2 Descriptive statistics

Figures 17 and 18 summarise the performance of the CURF, multiply imputed synthetic data based on a non-parametric hierarchical Bayes imputation model with Dirichlet Process precision parameter set to $\alpha = 19$, and Muralidhar and Sarathy’s perturbation method in estimating mean weekly income for groups defined by sex, ethnic group, educational achievement, sex by ethnic group and sex by educational achievement as listed in Table 1, as well as mean income for the full sample. The results reported in Figure 17 refer to imputation and perturbation models fitted to the IS 2003 data on the original scale, while Figure 18 reports results for log-scale imputation and perturbation modelling.

For the CURF, discrepancies of about $20 per week are apparent for the overall sample estimate (the rightmost point) and for several subgroups. The maximum discrepancy of $195 per week occurred for the second smallest group, (males, other ethnic combinations, n=46). However substantial discrepancies also occurred for the moderately sized groups with university qualifications, where discrepancies of $52 and $37 for per week were observed respectively for males (n=841) and females (n=915), leading to a discrepancy of $44.40 per week for the overall university qualified group. The university qualified groups in the IS2003 are likely to include some high earners and therefore it is plausible that these groups were more affected by top-coding of income and, possibly, record-deletion than other groups.

The maximum discrepancy between the real and multiply imputed synthetic weekly income estimates occurred for the female, other ethnic combinations (n=41) and female, Māori combinations groups (n=169) for which discrepancies of $36.69 and $37.12 were observed. Discrepancies declined with sample size and the discrepancy between the synthetic data and real data estimate for the overall sample was $3.34. Average income estimates under the perturbation method were identical to the actual estimates for all groups considered here, reflecting the mean-preserving property of the perturbation method for all groups defined by factors included in the perturbation models (all groups defined by two-way stratification in this case).

The second row of Figure 17 indicates that credible interval lengths for the CURF and synthetic data estimates were similar to those obtained from the real data. However perturbation method credible interval lengths varied somewhat from the real data interval lengths for several groups, reflecting the fact that although the perturbation method preserves overall variance it does not necessarily preserved sub-group variances.

The last row of Figure 17 presents the Karr measure of data utility, based on credibility interval overlap, for estimating mean income under the three confidentialising methods. Using this measure of data utility it appears that the synthetic data and perturbation methods produce estimates of mean income of high utility for all groups considered, whereas the CURF estimates for the smallest and largest groups have low utility. For the smallest groups, this is due to the large absolute discrepancies between CURF and IS2003 point estimates, The low data-utility for CURF estimates for larger groups reflects the fact that that the moderate absolute discrepancies of about $20 per week for these groups is large relative to the corresponding standard errors. For example, the standard error for the mean weekly income estimate for the overall IS2003 sample was $4.85. We note that the CURF was analysed using standard methods, without taking account of the confidentialising methods applied. Had some acknowledgement of the confidentialised nature of the data been made in the analysis, for example through the use of censored data methods for top-coded variables, (Lane, 2007), it is possible that the credible intervals for the CURF may have been longer than reported in Figure 17, and this may have led to a greater degree of credibility interval overlap and hence data utility according to the Karr measure.

The performance of multiply imputed synthetic data when the imputation model was fitted to the income and working hours data on the log-scale was similar to the performance for the original scale imputation (Figure 18). The major differences were for male Pacific (n=296) and female Māori groups (n=542). For the former group the discrepancy between synthetic data and real data mean income estimates increased from $1.28 to $22.12, resulting in decline in credibility overlap from 0.94 to 0.61. The latter value was the lowest data utility observed for synthetic data under the log-scale imputation model for any of the 36 groups considered. For the female Māori group the major change in moving
from original to log-scale imputation modelling was a substantial increase in the length of the credible interval with the relative credible interval length increasing from 1.1 to 5.7. This dramatic increase appears to have been due to a substantial over-estimation of the income standard deviation in this group, under the log-scale imputation model. In the real data the income standard deviation for the female Māori group was $282.66, whereas under the log scale imputation model the geometric mean synthetic income standard deviation for this group, obtained by exponentiating the average of the 100 log-transformed standard deviations (which better approximated the normality assumption underpinning the combining formulae) was, $456.55, whereas the under the imputation model fitted on the original scale, the geometric mean standard deviation was $287.91.

Comparing the third columns of Figure 17 and Figure 18, it can be seen that the perturbation method was adversely affected by moving to a log-scale for fitting the perturbation model. Substantial discrepancies between log-scale based perturbation and real data estimates of mean weakly income were apparent (see top right plot in Figure 18) with resulting decreases in data-utility. This reflects the fact that when the perturbation model is fitted on the log-scale the parameters of the log-scale regression and the means of log transformed variables are preserved, but mean values of the untransformed variables are not.
Figure 17

Comparison of CURF, MI: NPHB (\(\alpha = 19\)) and SBP: Estimating mean income for 36 groups.
Imputation/perturbation modelling on original scale.

CURF

MI: NPHB (\(\alpha = 19\))

perturbation

Comparison of CURF, MI: NPHB (\(\alpha = 19\)) and SBP: Estimating mean income for 36 groups.
Imputation/perturbation modelling on original scale.
Figure 18

Comparison of CURF, MI: NPHB (\(\alpha = 19\)) and SBP: Estimating mean income for 36 groups.

Imputation/perturbation modelling on log scale.

<table>
<thead>
<tr>
<th>CURF</th>
<th>MI: NPHB ((\alpha = 19))</th>
<th>Perturbation</th>
</tr>
</thead>
<tbody>
<tr>
<td>absolute mean difference</td>
<td>mean difference ($)</td>
<td>absolute mean difference</td>
</tr>
<tr>
<td>[Graphs showing comparisons]</td>
<td>[Graphs showing comparisons]</td>
<td>[Graphs showing comparisons]</td>
</tr>
<tr>
<td>relative ci length</td>
<td>relative ci length</td>
<td>relative ci length</td>
</tr>
<tr>
<td>[Graphs showing comparisons]</td>
<td>[Graphs showing comparisons]</td>
<td>[Graphs showing comparisons]</td>
</tr>
<tr>
<td>data utility</td>
<td>data utility</td>
<td>data utility</td>
</tr>
<tr>
<td>[Graphs showing comparisons]</td>
<td>[Graphs showing comparisons]</td>
<td>[Graphs showing comparisons]</td>
</tr>
</tbody>
</table>
The performance of the CURF, multiply imputed synthetic data, NPHB ($\alpha=19$) and perturbed data methods in estimating the median and lower and upper quartiles, under imputation and perturbation modelling of income and working hours on the original scale is illustrated in Figure 19. The corresponding results obtained under log scale imputation and perturbation models are shown in Figure 20.

With the exception of the three smallest groups, represented by the three leftmost points in the Figures the CURF and synthetic data methods produced estimates of income quartiles which were similar to those obtained from the real data. Nevertheless, for the three smallest groups (male, other ethnic combinations; female, other ethnic combinations; total other ethnic combinations) discrepancies were substantial, particularly for the CURF upper quartile estimates for which the discrepancy between the CURF and real data estimates for the male, other ethnic combinations group was approximately $240. In contrast to the CURF and synthetic data methods, income quartiles estimated from perturbed data were substantially different from the real data for the majority of groups considered and discrepancies appeared independent of sample size. Performance of the perturbation method was particularly poor for estimation of the upper quartile with discrepancies of $100$ or more for 35/36 and 20/36 groups under the original and log-scale perturbation models, respectively.
Figure 19

Comparison of CURF, MI: NPHB ($\alpha = 19$) and SBP: Estimating income quartiles
Imputation/perturbation modelling on original scale.

**CURF**

**MI: NPHB ($\alpha = 19$)**

**perturbation**

- Difference in medians
- Difference in LQ
- Difference in UQ

<table>
<thead>
<tr>
<th>Difference in Medians</th>
<th>Difference in LQ</th>
<th>Difference in UQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute difference ($)</td>
<td>Absolute difference ($)</td>
<td>Absolute difference ($)</td>
</tr>
<tr>
<td>$log_{10}(n)$</td>
<td>$log_{10}(n)$</td>
<td>$log_{10}(n)$</td>
</tr>
</tbody>
</table>
8.3 Log-linear regression models

We compare the performance of the CURF, multiply imputed synthetic data and perturbation methods for the same sequence of congenial and uncongenial scenarios considered in section 7.3 where now the concept of congeniality extends also to the model underlying the perturbation method. As discussed in section 8.1 the structures considered for the perturbation models matched the prior imputation model structures adopted for the parametric and non-parametric hierarchical models. The concept of congeniality does not apply to the CURF, because it was not created using model-based methods. The performance results for the CURF are therefore repeated, below, in Figures illustrating the performance of the multiple imputation and perturbation methods under alternative prior imputation / perturbation models.
8.3.1 Congenial Models

Figure 21 summarises the performance of the CURF, synthetic data based on an non-parametric hierarchical imputation model with $\alpha = 19$ and the sufficiency preserving perturbation (SBP) method in estimating a main effects log-linear regression for income on log-working hours and categorical predictors. Results for a log-linear regression model including an age by sex interaction, defined as the product of a linear age term and a binary sex indicator, in addition to log-working hours and categorical predictors is summarised in Figure 22. The prior imputation and perturbation models used for this scenario were specified on the log-scale for both weekly income and hours worked and included all two-way interactions between predictors (log working hours and categorical variables) but only included a linear age term in age related interactions. Under this scenario the analysis model structures on which Figures 21 and 22 are based are subsets of the prior imputation and perturbation model, and consequently the synthetic data and perturbation methods can be expected to perform well.

From Figures 21 and 22 it is clear that the perturbation methods achieved perfect performance in these congenial scenarios, as predicted by the theory of the perturbation method. The synthetic data method also achieved high data utilities for all parameters, despite some variability in relative credible interval length. Data utilities for the intercept and log-working hours parameters were less than 0.4 for the CURF, for both analysis models. This was due to small absolute point estimate discrepancies between CURF and real data estimates, which were nevertheless large relative to the real data standard errors. For example, for the main effects log-linear regression model the real data intercept was 3.77, whereas the CURF estimate was 3.81 and this small discrepancy was equivalent to 2.7 real data standard errors.
Figure 21

Comparison of CURF, MI and SBP: imputation / perturbation modelling on log scale. Estimating income main effects log-linear regression, working hours log transformed.

CURF MI: NPHB ($\alpha = 19$) perturbation

- Absolute Percentage Difference
- Absolute Percentage difference
- Absolute percentage difference

- Relative ci length

- Data utility

Comparison between CURF, MI and SBP using imputation or perturbation techniques on a log scale. Income main effects are estimated using log-linear regression with working hours log-transformed. The figure illustrates the comparison for CURF and MI with NPHB ($\alpha = 19$) perturbation.
8.3.2 An analysis model that is uncongenial due solely to the inclusion of non-linear components of the age effect in age by sex interactions

Figure 23 summarises the performance of the CURF, synthetic data and perturbation methods in estimating a log-linear regression model of income on log-working hours, categorical predictors and two-way interactions as described for Model 6 in Table 3. The prior imputation and perturbation model for this scenario were specified on the log-scale for both weekly income and working hours and the only difference between the prior imputation / perturbation and analysis model structures was that the analysis model included the full quadratic spline representation of age in the age by sex interaction term, whereas the prior imputation / perturbation model included only a linear age term in the age by sex interaction. This scenario represents a mild form of uncongeniality.

The results for this uncongenial scenario were somewhat mixed. In terms of the credible interval overlap based measure of data utility, the synthetic data method exhibits more consistently good performance than the other two methods with data utilities no less than 0.695. For the perturbation method a cluster of five parameters with data utility less than 0.2 can clearly be seen in Figure 23, and...
Confidentialising Microdata

a further 8 parameters had data utilities between 0.2 and 0.6. Nevertheless, under the perturbation method two-thirds of the 48 parameters had data utilities of 0.8 or more. Although this is less than the 91.7% of data utilities greater than 0.8 achieved by the synthetic data method, the proportion of parameters for which the data utility index was at least 0.9 were 60.4% (29/48) for both methods. Low data utilities for the perturbation method were clearly not confined to the age-sex interaction (which is represented by three parameters) and were attributable to point estimate discrepancies which are obscured in the plot of absolute percentage differences by the presence of a very large discrepancy for the ‘Other’ ethnicity by secondary qualifications interaction term.

The CURF is unaffected by the choice of prior imputation / perturbation model. The results for the CURF estimate of the model with interactions are summarised in Figure 23. The utility of CURF estimates of the intercept and log working hours parameters was low. However, data utilities of 0.8 or more were observed for 40 of the 48 parameters (83.3%) and although this was less than the corresponding proportion achieved by the synthetic data method (44/48 or 91.7%), the proportion of parameters with data utilities at least 0.9 was slightly higher for the CURF than for the other two methods (68.9% compared to 60.4%).
8.3.3 Models that are uncongenial due only to differences in the scale on which models are specified.
Prior imputation model specified on original scale; analysis model regresses log-income on untransformed working hours and categorical predictors.

To illustrate the impact of discrepancies in the scale on which prior imputation / perturbation and analysis models are specified we consider here a scenario in which the prior imputation and perturbation models were specified on the original untransformed scale for both weekly income and hours worked, whereas the analysis model was specified as a log-linear regression of income against working hours (untransformed) and categorical predictors. Thus, whereas the prior imputation / perturbation model assumed a linear relationship between income and hours worked the analysis model for this scenario assumes a linear relationship between log income and actual (ie untransformed working hours). The performance of the CURF, synthetic data and perturbation methods in estimating this model summarised in Figure 24.
The performance of the perturbation method under this uncongenial scenario was very poor with consistently high relative credible interval lengths and data utilities close to zero for several parameters. Data utilities for the perturbation methods did not exceed 0.7 for any of the parameters. Although the synthetic data method improved on the perturbation method, data utilities for the synthetic data estimates were greater than 0.8 for only 3 of the 15 parameters and for the working hours parameter the data utility for synthetic data estimate was only 0.002 (second leftmost point on the plot). These results are worse than was observed for the corresponding scenario in the exercise reported on in section 7, in which the CURF served as the real dataset (see Figure 8). With the exception of the working hours parameter, data utility for the CURF parameter estimates was at least 0.73. The data utility for the CURF estimate of the working hours parameter was 0.23.
Prior imputation / perturbation model specified on log- scale; analysis model regresses log-income on untransformed working hours and categorical predictors.

We now consider an ungenial scenario in which the prior imputation / perturbation and analysis model specifications disagree only on the representation of the working hours variable. This is in contrast to the scenario explored above where prior imputation and analysis models differed only with respect to the scale on which the outcome variable (income) was modelled. In the scenario considered here the prior imputation / perturbation model is specified on the log-scale for both weekly income and hours worked, whereas the analysis model regresses log-income against working hours (untransformed) and categorical predictors. The performance of the CURF, synthetic data method and perturbation method are summarised in Figure 25.

For this scenario the performance of the perturbation and synthetic methods was similar with data utility exceeding 0.8 for 11 and 12 of the 15 main effect model parameters for the perturbation and synthetic data methods respectively. The main difference between the performance of the perturbation and synthetic data methods was with respect to the 'other' ethnic group indicator for
which the data utility index was 0.60 for the perturbation method and 0.81 for the synthetic data method. The difference appears related to a combination of a smaller point estimate discrepancy under the synthetic data method (3.2%, equivalent to 1.2 real data standard errors for synthetic data, compared to 5.2% equivalent to 1.7 real standard errors for the perturbation method), combined with some inflation of synthetic data credible interval length relative to the real data interval length (relative credible interval lengths were 1.34 and 1.00 for synthetic and perturbed data respectively). This resulted in greater credible interval overlap under for the synthetic data method. Despite the overall good performance of the synthetic data and perturbation methods in this uncongenial scenario, both methods produced data utilities close to zero for the intercept, working hours and sex parameters. Although absolute point estimate discrepancies were not large, the low data utilities nevertheless indicate that the mis-modelling of the log-linear relationship between income and working hours assumed in the analysis model by the linear relationship between log income and log working hours specified for the prior imputation / perturbation model limits the ability of synthetic or perturbed data to reproduce the log-linear relationship estimated by the real data.

Since the CURF is unaffected by the choice of prior imputation model its performance in estimating the regression of log-income on untransformed working hours and categorical predictors is identical to those reported in Figure 24 for the preceding scenario. The CURF estimates for this analysis model are superior to those obtained by the synthetic or perturbed data when the prior imputation / perturbation model is uncongenial.
Prior imputation model specified on original scale; analysis model regresses log income against log working hours and categorical predictors.

We turn now to the scenario in which the scale on which both income and working hours are analysed differs between the prior imputation / perturbation model. The performance of the CURF, synthetic data and perturbation methods in estimating a main effects regression in which log income is regressed against log working hours and categorical predictors when the prior imputation / perturbation model is mis-specified as linear regression of untransformed income against untransformed working hours and categorical predictors is summarised in Figure 26. The performance of the perturbation method in this scenario was poor, with considerable inflation of credible interval length, relative to the real data intervals. Data utilities for nine of the 15 model parameters were less than $1 \times 10^{-4}$ and did not exceed 0.70 for any parameter. The performance of the synthetic data was better than the perturbation method for this uncongenial scenario, however the data utilities for the synthetic data estimates of the intercept and log-working hours parameters were close to zero and data utilities greater than 0.8 were achieved for only three parameters (sex, secondary qualifications and one component of the age quadratic spline). The performance of the CURF in estimating the regression of
log income on log working hours and categorical predictors was superior to that achieved by the synthetic data or perturbed data in this uncongenial scenario. With the exception of the intercept and log-working hours parameters for which the CURF produced data utilities of 0.21 and 0.36, data utilities exceeded 0.7 for all parameters and were at least 0.8 for 10 of the 15 model parameters.
8.3.4 Models that are uncongenial due to differences in the scale used for model specification and inclusion of non-linear components of the age-sex interaction term

We now consider the joint impact of misuse specification of the scale on which income and working hours are modelled and exclusion of interaction terms from the prior imputation/ perturbation model. To investigate this issue we considered analysis models specified as log-linear regressions of income on working hours and categorical predictors, including two-way interactions, with working hours represented either on the log transformed scale or on the original scale, as described for models 6 and 9 in Table 3. We considered prior imputation/ perturbation models which omitted non-linear components of the quadratic spline representation for age from the age by sex interaction and also differed from the analysis models in the assumed scale on which income and working hours were represented. The sequence of scenarios considered parallels that presented in section 7.3.4. The performance of the CURF, synthetic data and perturbed data for these scenarios is summarised in Figures 27 to 29 and a summary of the credibility interval overlap based data utility indices is presented in Table 6.
The pattern of results for the three confidentialising methods is similar to that reported in section 8.3.3. For both scenarios in which the prior imputation model was specified on the original scale for both income and hours worked (see Figure 27 and Figure 29) the performance of the perturbation method is poor with data utility less than 0.5 for the majority of parameters and less than 0.2 for approximately 30% of model parameters (15/48) when untransformed working hours is used as a regressor in the analysis model and 14/48 when log transformed working hours is used in the analysis model). In these two scenarios (labeled A and C in Table 6) the synthetic data method performed better than the perturbation method but nevertheless achieved data utilities of at least 0.8 for only 25% and 18.8% of parameters in the two scenarios. The performance of both the synthetic data and perturbation methods was considerably better for the scenario in which the prior imputation / perturbation model was specified for log income and log working hours and the analysis model included working hours on its original scale (scenario B in Table 6). For this uncongenial scenario data utilities of at least 0.8 were achieved for 41 of the 48 (85.4%) parameters using the synthetic data method and for 36 of the 48 parameters (75%) using the perturbation methods. Nevertheless data utilities remained low for some parameters for both synthetic data and perturbed data, under this scenario. The performance of the CURF in estimating both the two-way interaction analysis model including untransformed working hours and the corresponding model including log-transformed working hours was superior to the performance of the synthetic data and perturbation methods under these uncongenial scenarios.
Figure 27

Comparison of CURF, MI and SBP: imputation modelling on original scale.
Estimating income log-linear regression with interactions, working hours untransformed.

CURF

MI: NPHB \(\alpha = 19\) perturbation

Absolute Percentage Difference

Absolute Percentage difference

Absolute percentage difference

Relative ci length

Relative ci length

Relative ci length

Data utility

Data utility

Data utility

Confidentialising Microdata
Figure 28


<table>
<thead>
<tr>
<th>CURF</th>
<th>MI: NPHB (α = 19)</th>
<th>perturbation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute Percentage Difference</td>
<td>Absolute Percentage difference</td>
<td>Absolute percentage difference</td>
</tr>
<tr>
<td>% difference</td>
<td>% difference</td>
<td>% difference</td>
</tr>
<tr>
<td>log standardised standard error</td>
<td>log standardised standard error</td>
<td>log standardised standard error</td>
</tr>
<tr>
<td>Relative ci length</td>
<td>Relative ci length</td>
<td>Relative ci length</td>
</tr>
<tr>
<td>relative ci length</td>
<td>relative ci length</td>
<td>relative ci length</td>
</tr>
<tr>
<td>log standardised standard error</td>
<td>log standardised standard error</td>
<td>log standardised standard error</td>
</tr>
<tr>
<td>Data utility</td>
<td>Data utility</td>
<td>Data utility</td>
</tr>
<tr>
<td>Karr index</td>
<td>Karr index</td>
<td>Karr index</td>
</tr>
<tr>
<td>log standardised standard error</td>
<td>log standardised standard error</td>
<td>log standardised standard error</td>
</tr>
</tbody>
</table>
Figure 29

Comparison of CURF, MI and SBP: imputation modelling on original scale.
Estimating income log-linear regression with interactions, working hours log transformed.

CURF

MI: NPHB ($\alpha = 19$)
Table 6

Summary statistics for credible interval overlap based data utility for the CURF, synthetic and perturbed data under three uncongenial scenarios

<table>
<thead>
<tr>
<th>scenario</th>
<th>statistic</th>
<th>CURF</th>
<th>Synthetic data&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Perturbed data</th>
</tr>
</thead>
<tbody>
<tr>
<td>A&lt;sup&gt;1&lt;/sup&gt;</td>
<td>minimum data utility</td>
<td>0.50</td>
<td>0.15</td>
<td>0.00</td>
</tr>
<tr>
<td>A</td>
<td>% data utilities &lt; 0.2</td>
<td>0.0</td>
<td>2.1</td>
<td>31.2</td>
</tr>
<tr>
<td>A</td>
<td>% data utilities ≥ 0.5</td>
<td>100.0</td>
<td>77.1</td>
<td>64.6</td>
</tr>
<tr>
<td>A</td>
<td>% data utilities ≥ 0.8</td>
<td>93.8</td>
<td>25.0</td>
<td>0.0</td>
</tr>
<tr>
<td>A</td>
<td>% data utilities ≥ 0.9</td>
<td>81.3</td>
<td>10.4</td>
<td>0.0</td>
</tr>
<tr>
<td>B&lt;sup&gt;2&lt;/sup&gt;</td>
<td>minimum data utility</td>
<td>0.50</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>B</td>
<td>% data utilities &lt; 0.2</td>
<td>0.0</td>
<td>6.3</td>
<td>14.6</td>
</tr>
<tr>
<td>B</td>
<td>% data utilities ≥ 0.5</td>
<td>100.0</td>
<td>93.8</td>
<td>85.4</td>
</tr>
<tr>
<td>B</td>
<td>% data utilities ≥ 0.8</td>
<td>93.8</td>
<td>85.4</td>
<td>75.0</td>
</tr>
<tr>
<td>C&lt;sup&gt;3&lt;/sup&gt;</td>
<td>minimum data utility</td>
<td>0.11</td>
<td>0.22</td>
<td>0.00</td>
</tr>
<tr>
<td>C</td>
<td>% data utilities &lt; 0.2</td>
<td>4.2</td>
<td>0.0</td>
<td>29.1</td>
</tr>
<tr>
<td>C</td>
<td>% data utilities ≥ 0.5</td>
<td>93.8</td>
<td>87.5</td>
<td>64.6</td>
</tr>
<tr>
<td>C</td>
<td>% data utilities ≥ 0.8</td>
<td>83.3</td>
<td>18.8</td>
<td>0.0</td>
</tr>
<tr>
<td>C</td>
<td>% data utilities ≥ 0.9</td>
<td>68.8</td>
<td>6.3</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<sup>1</sup> Prior imputation / perturbation model specified on original scale for both income and hours worked; Analysis model specified as log-linear regression for income against untransformed working hours, categorical variables and interactions including non-linear components of the age effect in an age by sex interaction.

<sup>2</sup> Prior imputation / perturbation model specified on log scale for both income and hours worked; Analysis model specified as log-linear regression for income against untransformed working hours, categorical variables and interactions including non-linear components of the age effect in an age by sex interaction.

<sup>3</sup> Prior imputation / perturbation model specified on original scale for both income and hours worked; Analysis model specified as log-linear regression for income against log-transformed working hours, categorical variables and interactions including non-linear components of the age effect in an age by sex interaction.

<sup>a</sup> Synthetic data based on a non-parametric hierarchical imputation model with Dirichlet Process precision parameter set to $\alpha = 19$. The centering model was a hierarchical multivariate normal regression model.
9 Discussion

In this report we have presented a new approach for formulating imputation models for multiply imputed synthetic data which produces synthetic datasets comprising a mix of numerical and categorical variables. The model builds on previous work for categorical data which emphasised the use of hierarchical models to provide some protection against mis-specification of the prior imputation model relative to the model assumed by the analyst. The new model builds on the hierarchical categorical data model by modelling numerical variables conditionally on the categorical variables, embedding the model for numerical data in a hierarchical structure, and using the hierarchical model structure as the centering model in a non-parametric Bayesian set-up in which the data is modelled using a Dirichlet Process prior for the distribution from which the observed data are assumed to have been sampled. The posterior predictive distribution for the numerical data under this set-up is a generalised Polya urn which produces synthetic data comprising a mix of values generated from the centering hierarchical model and values drawn from the original data.

The re-introduction of actual data values into the synthetic datasets provides additional protection against mis-specification of the prior imputation model over that provided by hierarchical models. Such additional protection is desirable for numerical data because hierarchical models generally embed only the model for the mean in a probability model and while this improves estimation of the mean it does not guarantee good estimation for other features of the data distribution. By treating a conventional hierarchical model, such as a hierarchical multivariate normal regression model, as the expectation of the data distribution, instead of assuming that the hierarchical model defines the data distribution, uncertainty about the form of the data distribution is reflected in the generation of synthetic data. The Dirichlet Process precision parameter represents this uncertainty. Large values of the precision parameter lead to synthetic data values being drawn relatively frequently from the hierarchical model, whereas low values of the precision parameter lead to observed data records being sampled relatively frequently. In fact, in the limit as the precision parameter approaches zero the imputation model proposed in this report for numerical data reduces to a series of cell-specific Bayesian bootstrap imputations.

Our implementation of non-parametric Bayesian hierarchical imputation models for numerical data was limited to models with a hierarchical multivariate normal regression model as the centering distribution. This simplified implementation, however the concept of embedding a hierarchical model in non-parametric Bayesian model could in principle be applied to any hierarchical (or non-hierarchical) imputation model including multivariate models specified as a sequence of conditional univariate distributions. Moving away from the multivariate normal may permit additional flexibility in modelling, for example by avoiding the linearity assumptions for associations between component variables that is inherent in the multivariate normal model.

The notion of protecting synthetic data imputations against mis-specification of the prior imputation model is important because a statistical agency releasing data cannot control the analyses actually undertaken by external users and, ideally, an analyst would obtain trustworthy results from their analysis even when their model differs from the imputer’s prior model in some respects. While achieving such robustness for all possible analyses is a daunting task it nevertheless seems reasonable that, other things being equal, a statistical agency should prefer imputation methods that generate synthetic data which admit good inferential performance across a range of analyses, including those that are not congenial to the imputer’s prior model (Meng, 1994).

Results reported in section 7 suggest the non-parametric hierarchical imputation model offers worthwhile improvements in robustness to prior imputation model specification compared to fully parametric imputation models. For example, Figures 1 and 2 show clear reductions in differences between synthetic data and real data point estimates of means and improvements in credible interval overlap, under non-parametric compared to fully parametric hierarchical imputation modelling. Similarly, in the uncongenial scenarios explored in 7.3, in which there is some degree of incompatibility between the prior imputation and analysis models, non-parametric hierarchical imputation models out-performed fully parametric hierarchical models for all scenarios except the scenario in which the prior imputation model specified a linear relationship between log income and log hours worked, whereas the analysis model assumed a linear relationship between log income and untransformed working hours. Under the latter scenario, both fully-parametric and non-parametric hierarchical imputations models produced reasonable synthetic data estimates for the majority of parameters in a main effects linear regression of log income against hours worked and several categorical predictors but exhibited very poor data utilities for three parameters (intercept, working hours and sex). It seems that the linearity assumptions inherent in the adoption of multivariate normal
Confidentialising Microdata

model for numerical variables still impact on synthetic data even when the multivariate model is embedded in a non-parametric hierarchical model.

Figure 13 suggests some improvement when the Dirichlet Process parameter is set to nine rather than nineteen, suggesting that further reductions may be required to overcome the impact of prior imputation model mis-specification in this case. Interestingly, under an alternative uncongenial scenario in which a linear relationship between income and working hours was specified in the imputation model but a linear relationship between log income and log–working hours was assumed by the analysis model the non-parametric hierarchical imputation model performed well (see Figure 14) and clearly improved on the fully parametric hierarchical model. It is unclear why the impacts of these two types prior of imputation model mis-specification differ.

While non-parametric hierarchical imputation models appear to offer advantages over fully parametric hierarchical models, results reported in section 8 suggest that multiply imputed synthetic data based on non-parametric hierarchical imputation models does not, in general, improve on the CURF in terms of quality of inferences obtained. Nevertheless, the synthetic data methodology did appear to offer some advantage for estimating means for small and large subgroups (see Figures 17 and 18). We note that in regression models the CURF often produced estimates of the effect of working hours with low data utility, as measured by credibility interval overlap. This may be a result of top-coding distorting the true relationship between income and working hours, although absolute differences in point estimates computed from the CURF and from original IS2003 data were generally small.

Sufficiency based perturbation proved extremely sensitive to mis-specification of the perturbation model with model parameter estimates obtained from perturbed data being close to real data estimates only for congenial analysis and perturbation models. Although sufficiency based perturbation provided perfect estimation of means when the perturbation model was specified on the original scale, neither original nor log-scale specification of the perturbation model provided accurate estimates of income quartiles. The sensitivity of the sufficiency based perturbation method to specification of the perturbation model would appear to limit the utility of this method as the basis of a general data release system.

The approach to constructing multiple imputation models for numerical data presented in this report could be improved and extended in a number of different directions. Firstly the reliance on multivariate normality for the hierarchical centering model could be relaxed and this may give more flexibility in modelling specific components of the joint distribution of the numerical variables. As previously noted, in some uncongenial scenarios the linearity assumptions inherent in the multivariate normal model appeared to impact on synthetic data even when the hierarchical normal model was embedded in a non-parametric Bayesian model. Replacing the multivariate normal model with a sequence of conditional univariate regression models, with some of these regressions using flexible regression forms may allow an imputation model to adapt more closely to the data. Fitting a sequence of models is also the most appealing strategy for extending the non-parametric hierarchical imputation model to higher dimensional problems. That is, if a p-dimensional set of variables falls naturally into k groups they can be modelled by a sequence of k models with all but the first block of variables being modelled conditionally on the preceding sets. Such a strategy would also allow a statistical agency to release synthetic data sequentially, as each imputation model is developed to a satisfactory standard. This would permit an agency to respond to requests for additional variables subsequent to the release of synthetic dataset. Moreover, such sequentially released synthetic datasets would be consistent since values of newly released variables would be generated conditionally on the synthetic values already generated for the previously released variables. The generated values for the newly synthesised variables would simply be added as additional columns to the existing synthetic datasets.

Another direction for extension concerns the non-parametric Bayesian aspects of the imputation model. There may be a case for treating the Dirichlet Process precision parameter as an unknown to be estimated from the data rather than holding it fixed as in our current implementation. This could, in a sense, help optimise the imputation model because poor fit of the hierarchical centering model would result in a posterior for the Dirichlet Process precision parameter concentrated on low values. This would have the effect of automatically compensating for a poor choice of centering model by ensuring that the original data was sampled relatively frequently during the creation of synthetic data. On the other hand, there is also an argument for fixing the precision parameter since this permits an agency to control, indirectly, the frequency with which real data records are sampled. A more fundamental change to the non-parametric hierarchical model would be to move the non-parametric Bayesian model to the second stage of the hierarchical model. In such a set-up the first stage of the model would be a standard parametric distribution, such as a multivariate normal, but the current multivariate normal model for the cell means would be replaced by a non-parametric Bayesian model centered on the multivariate normal. This would result in fully synthetic data because at the final stage
of imputation synthetic data would be drawn from the parametric data model. However, the adoption of a non-parametric Bayesian model for the expected cell means should greatly improve the fit of the overall model compared to a fully parametric hierarchical model, so good inferential performance for synthetic data generated from such a model could be expected. Although strategies for fitting hierarchical models with non-parametric models replacing standard parametric forms at the second stage of the model have been developed (Ohlssen, 2007; Ishwaran & James, 2001) our initial exploration of this approach indicated that computing times are likely to be considerably longer than for the Gibbs sampler we developed to implement our model.

Application of non-parametric hierarchical imputation models to the creation of synthetic versions of data types other than cross-sectional survey data would be a worthwhile direction for further research. For example, the problem of creating synthetic longitudinal data has received some attention (Abowd and Woodcock, 2001) and hierarchical models are an appealing choice for imputation modelling in the longitudinal setting because they provide a natural framework for accommodating the correlated nature of repeated data from the same individual. In a hierarchical model for longitudinal data the repeated assessments for each individual form the lowest level in the model and are assumed to be drawn from a model which is dependent on a parameter specific to the individual concerned. These individual-specific parameters (commonly referred to as random effects) are linked in a second level model. A non-parametric Bayesian model could be considered for this second level model.

In conclusion, we have presented a new approach for constructing imputation models for mixed numerical and categorical datasets which combines hierarchical models with non-parametric Bayesian models. This approach offers improvements over fully parametric hierarchical models in the robustness of synthetic data to mis-specification of the prior imputation model, relative to the analysis model. The proposed new methodology produces synthetic data which is a mix of records drawn from the original dataset and model-generated records but the records drawn from the original data are drawn in a random manner and an external user would be unable to distinguish model-generated from real data records. A preliminary investigation of disclosure risks associated with our proposed new imputation method suggests that the risks associated with introducing some real data records into synthetic datasets may be tolerable. Our current implementation of non-parametric hierarchical imputation models for numerical data is centered on a hierarchical multivariate normal regression model however other more flexible model structures could be used to center the non-parametric model. Synthetic data based on the non-parametric, hierarchical multivariate normal imputation model structure did not, in general, improve on Statistics New Zealand’s current CURF methodology, in terms of the utility of inferences obtained from the respective datasets. However, our synthetic data method offered a modest advantage over the CURF for some descriptive statistics in small and large subgroups. Both the CURF and the proposed synthetic data method seem preferable to sufficiency based perturbation which is extremely sensitive to the choice of perturbation model.
References


Confidentialising Microdata


