

Multiply Imputed Synthetic Data Files

Patrick Graham¹ and Richard Penny²

¹Senior Research Fellow,
Department of Public Health and General Practice,
University of Otago, Christchurch.

²Senior Methodologist, Statistics New Zealand

Abstract

In this paper we review the Bayesian logic which underpins a proposal of Rubin to apply multiple imputation to the problem of constructing synthetic versions of survey datasets which official statistics agencies can release to researchers without compromising the confidentiality of the survey responses. We investigate the application of the multiple imputation paradigm to constructing synthetic versions of a small dimensional but real dataset. We emphasise the use of hierarchical Bayesian imputation models in order to reduce the dependence of synthetic datasets on specific structural model forms. In our example, analyses of synthetic data generated under hierarchical Bayesian models exhibit more robustness to the imputation model than do analyses of synthetic data based on non-hierarchical versions of the same imputation models. However, so long as the analytical model is simpler than the imputation model, the impact of the imputation model on inference appears slight. The confidentiality aspects of synthetic data are considered in the context of our example and we conclude that disclosure risks are no greater, and probably less, than those associated with the established confidentialising technique of random rounding.

Keywords

Bayesian modelling; Synthetic data files; Multiple imputation; Statistical disclosure limitation;

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 authors, 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. The authors thank T.E. Raghunathan for helpful comments on aspects of the theory of multiply imputed synthetic datafiles and Alistair Gray, Jim Young and Jerry Reiter for comments on an earlier version of this paper.

1 Introduction

An Official Statistical System Agency (OSSA) collects and stores data collected from respondents in what we will term a Unit Record File (URF). In the majority of cases the OSSA will try to ensure that no-one, other than those entitled to have access to the data, can see, or derive from outputs from the data, the responses of individual respondents. However the data are collected to be used for various purposes, be they administrative, legal or research. A major issue the OSSA faces is providing outputs from the URF, while preserving the confidentiality of the information in each unit record (Duncan et al. 1993, Willenborg & De Waal 1996, Doyle et al. 2001).

While for many of these outputs the data can be kept in-house by the OSSA, there is an increasing need to enable people from outside the OSSA to have access to parts of the data. This is particularly crucial for research work, one of the main drivers being the desire for more evidence based policy focussed on particular subgroups.

For many years researchers worked with tables produced by the OSSA from the URF. To preserve the confidentiality of the responses used to generate the tables many techniques have been developed to prevent a researcher from deducing individual responses from the numbers in the table (e.g. random rounding and cell suppression). However, in the last few decades greater computing power combined with larger amounts and more detailed data being collected has led researchers to demand access to the individual responses. We note that when only a small number of categorical variables are required for analysis there is no essential difference between a URF and a tabular representation of the URF. The distinction is therefore not some much between tables and URFs as between data which can be conveniently represented via a multiway table and data which is sufficiently multidimensional to make a URF a more convenient representation. While names, addresses and identifications numbers (e.g. SSN, IRD number) are generally not stored in a URF, in very many cases it would possible to identify individual respondents due to their unique combination of responses to the variables collected. It is obvious that if enough variables with enough details are in the URF most, if not all, records will end up being unique in the file. As a result of the need for disaggregated data there has been considerable research investigating approaches that would allow researchers to work with data held by an OSSA or other data collection agency, yet prevent a confidentiality breach that would allow identification of one or more respondents.

A common approach to giving researchers access to data is to bring them into the OSSA. As such they are required to conform to the same restrictions as those within the OSSA, and subject to the same legal penalties if they breach respondent confidentiality. This can be inconvenient for researchers based at locations remote from the OSSA

facilities. In any case the researcher will be working in an unfamiliar environment and accessing data at the convenience of the OSSA. A further difficulty is that the data is unavailable to those who may wish to review and assess the analysis done by the researcher, unless they too can gain this access.

Since the 60's some sample URFs have been available to researchers. That is, either the URF comes from a sample survey, or else a sample of records are selected from the data file in the OSSA. While the sample can itself be drawn from a sample, for the purposes of this paper we will assume that a sample is drawn from the total population.

A sample has a large amount of protection as the researcher is usually not aware who has, and has not, been sampled. However there is the risk of that a unique individual may be in the sample (e.g. the Prime Minister). In this case the individual is clearly known to be unique in the source population from which the sample was drawn. A more interesting problem is when one has an unusual individual (e.g. 27 year old married plumber with a PhD). Is this person actually unique, and thus identifiable? That is, is it your neighbour who is a 27 year old plumber with a PhD?

It is clear that URF are an important and increasingly necessary output required from OSSAs so the issue is "how can they be released?" That is, is it possible to produce a Confidentialised Unit Record File (CURF) that will provide a suitable level of protection for the respondent yet provide researchers with similar inferences to those they would obtain from the original URF. If this is feasible then researchers can work with the CURF and be assured the models they fit, and the analytical results from those models, are consistent with those they would obtain if they worked with the URF, and yet respondent confidentiality is preserved.

2 Approaches to constructing confidentialised data files

2.1 Confidentialised unit record files via ad-hoc adjustments

2.1.1 Adjustment procedures

Given that identifiable individuals will be those who have a unique combination of responses a conceptually simple approach is to identify those unique individuals and make changes in the level of detail that make them unique. A common approach is to remove identifying variables. Obvious examples would be name and address, but identifying variables such as date of birth or geographical location are often made unavailable. If a variable is necessary for the research the level of detail in that variable may be limited. For example date of birth is often used to determine age so while the date of birth will not be released a respondent's age may be. However it may be

necessary to release age in 5 year age group bands, rather than as individual ages. Extreme values often make a respondent readily identifiable so these extremes could be suppressed which may mean particular subpopulations are missing. Another approach is to modify the data, say by putting all respondents earning more than \$70,000 in a class \$70,000+ rather than providing their exact income. This removes this aspect of their uniqueness and thus their identifiability. There are also many other ways to alter the unit record to prevent statistical disclosures in URF (see Chapter 4 Willenborg & de Waal 1996).

The difficulties with trying to identify and modify unique and identifiable individuals in this way are many. First, it is very labour intensive and many judgements are required as to how to modify the data so that confidentiality is preserved while minimising the loss of detail in unit records. Then there is the issue of making these modifications in such a way that the interesting parts of the data are not so heavily modified that informative analysis is impossible. Another difficulty with this approach is that it doesn't have a readily repeatable set of principles that ensure consistency between how each URF will be confidentialised, nor does it provide guidance to the user on how analytical methods should be modified to reflect the confidentialising process.

A related issue is that it is difficult for the OSSA to know before commencing work on a CURF how much work will be required and whether the end result will be analytically useful. It is a serious waste of resources to create a CURF that turns out to be unusable.

2.1.2 Identifying variables to be adjusted

Given the statistical tools used to analyse the data it seems logical to apply statistical tools to assess the overall disclosure risk within an URF, and use these tools to assist in the identification of those variables, or particular characteristics of the variables that lead to disclosure risk. The OSSA could identify before it even attempts to create a CURF whether it is a feasible exercise and if so which variables need to be adjusted to protect confidentiality.

Suppose there are K distinct combinations of characteristics in a population. If F_i denotes the number of individuals in the population with the i^{th} set of characteristics, and f_i denotes the corresponding number in the sample, then for any given combination of characteristics i , with observed sample frequency equal to one, disclosure risk is simply $\Pr(F_i = 1 | f_i = 1)$. If this probability is one then releasing the observed URF without modification would constitute a clear confidentiality breach. However the OSSA may decide to set the acceptable probability that defines a confidentiality breach to be less than one. In any case the conditional distributions $p(F_i | f_i)$ are, in principle, required for all characteristic sets i .

Initially models of the conditional population frequency distributions were based on

superpopulation models of the form

$$\begin{aligned} F_i | \pi_i &\sim \text{Poisson}(N\pi_i) \\ f_i | \pi_i &\sim \text{Poisson}(n\pi_i) \\ n\pi_i &\sim \text{Gamma}(\alpha, \beta) \end{aligned}$$

where N denotes the population size, n the sample size and the π_i are the cell probabilities for the K cells (i.e. the K possible combinations of variable values). Bethlehem et al. (1990), Skinner et al. (1994) and others show that this model implies.

$$\Pr(\text{population unique}) = (1 + N\beta)^{-(1+\alpha)}$$

Fienberg & Makov (1998) note that this is a frequency of frequencies approach. If this model fits one could identify URFs, or parts of URFs, with major confidentiality problems and not make them available to researchers except within the OSSA. Less risky URFs would merit some effort to identify and modify the risky unit records.

Unfortunately when this theory is applied to realistic data it does not fit very well (Bethlehem et al. (1990)). Polletini and Stander (2004, section 5), have proposed a hierarchical Dirichlet-Multinomial model which appears more plausible than the Gamma-Poisson model given above. In contrast to the Gamma-Poisson model, the Dirichlet-Multinomial proposed by Polletini and Stander directly links population and sample frequencies. Building on work by Samuels (1998), Fienberg and Makov (2001) have developed an interesting approach using genetic-inspired urn models. However, much recent research has headed to a more structured approach to modelling, for example by using log-linear and logistic models (Fienberg & Makov 1998, Skinner & Holmes 1998). By fitting a structured model to the data it is hoped that this will account for the interactions between the variables and truly measure the disclosure risk in the URF.

2.2 Multiply imputed synthetic data.

Rubin (1993) proposed a different approach to releasing the statistical information contained in a URF, or more generally, an observed dataset (including tabular files) based on the ideas of multiple imputation (Rubin 1987). Imputation has been used for many years to provide a unit record when a respondent has not provided a response when surveyed. As the OSSA cannot be sure that the imputed unit record is the same as the response that would have been provided by the respondent, some extra uncertainty is added to any analysis of a URF containing imputed unit records. The problem multiple imputation was designed to solve was how this extra uncertainty could be estimated.

In very simple terms under multiple imputation one imputes for nonresponse more

than once (assuming that one has stochastic rather than deterministic imputation). Thus one ends up with several, URFs of the same size. Any analysis is done separately on each of the imputed URFs and the variation in the outputs from these analyses will be due to the uncertainty of the imputation. For conventional missing data problems Rubin (1987) suggested approximately five imputations would suffice but notes that when the fraction of missing information is large a larger number of imputations may be required.

Given multiple imputation is designed to impute unit records where there is no response supplied, Rubin's (1993) proposal for synthetic data was to multiply impute records for members of the population not included in the survey sample and to release a sample from each "completed" URF. Clearly the imputed unit records for respondents not in the survey are based on the information from those that were in the survey and, as outlined in the following section, the entire imputation process follows the logic of Bayesian predictive inference.

Under the multiple imputation (MI) paradigm the researcher works with several synthetic unit record files (SURF) and combines analyses from each SURF to obtain a single inference from the multiple SURFs. In each SURF a record may either be a record for someone who responded, or an imputed (i.e. synthetic) record. In typical situations involving small sampling fractions the proportion of SURF records corresponding to observed records would be low.

Fienberg and Makov (1998) used the multiple imputation (MI) approach to resolve the issue discussed in the previous section of modelling the probability distribution of cell counts by generating a SURF to estimate the probabilities for a disclosure risk. Interestingly they did not suggest using the SURF itself for analysis purposes.

Raghunathan et al (2003) and Reiter (2002) have shown that SURFs generated under the MI paradigm can produce inference-valid results for under simple random sampling and for some complex designs, although the validity of MI-based synthetic data has not yet been tested on the full range of design complexity employed by official statistics agencies. Given that all an OSSA needs to do to preserve a respondent's confidentiality is to make the researcher unsure that the unit record in the SURF corresponds to a real record, Kennickell (1997), Abowd and Woodcock (2001) and Liu and Little (2002) have explored a variant of Rubin's idea involving the release of a URF with some of the unit records replaced by imputed data.

In this paper we background the Bayesian logic underpinning a minor variant of Rubin's original proposal, also mentioned by Rubin (1993) and Raghunathan et al (2003), in which all records in the SURFs are synthetic and are generated via draws from the posterior predictive distribution for a new sample from the population, given the observed data. This approach is similar to that outlined in Reiter (2005). We also illustrate the potential benefits of hierarchical Bayesian (HB) approaches to develop-

ing imputation models. To our knowledge this represents the first application of HB methods to the development of multiply imputed synthetic datasets.

3 Bayesian predictive logic and synthetic data files

3.1 Bayesian finite population inference

The logic underpinning the proposal of Rubin and colleagues to base construction of synthetic files on a multiple imputation (MI) paradigm is best understood from the perspective of Bayesian finite population inference. Consequently, we present below heuristic arguments justifying the MI approach, based on Bayesian predictive inference, firstly in the case of sample from a finite population and secondly, in the case of a census of a population (section 3.5). The presentation below is a formalisation and synthesis of ideas in papers by Rubin (1993), Reiter (2002), and Raghunathan, Reiter and Rubin (2003) and emphasises the underlying Bayesian logic of the MI proposal.

Firstly, let $Y^{pop} = (Y^{obs}, Y^{mis})$ denote the $N \times p$ data matrix for a finite population of size N , where Y^{obs} denotes the observed data matrix for n randomly selected individuals and Y^{mis} denotes the unobserved data matrix for the remaining $N - n$, individuals not included in the study sample. Further, we suppose the object of inferential interest is denoted $Q = q(Y^{pop})$. The quantity Q may be a simple scalar statistic, such as the average value of some variable in the population, a vector of such quantities or more complicated entities such as sets of cross-classified margins of complex contingency tables. Potentially, Q may also be a parameter of some statistical model such as linear or logistic regression. This would be the case if given the full population data an analyst would fit a model to the full population and focus on the interpretation of model parameters or deterministic functions of these parameters, quoting point and interval estimates derived from the model. Technically, this moves the inferential problem from finite population to infinite super-population inference, that is to problems where the object of interest is a superpopulation quantity. Nevertheless, as discussed in section 3.5, below, the general imputation framework developed for the finite population case can be applied to super-population inference or to inference for model parameters.

Given the data for a sample from some population, the Bayesian approach to finite population inference is to base inference on the posterior distribution of the unobserved responses of the $N - n$ individuals not included in the sample. That is, Bayesian inferences follow from $p(Y^{mis}|Y^{obs})$. Formally, the posterior distribution for a finite

population quantity of interest is

$$\begin{aligned} p(Q|Y^{obs}) &= \int p(Q|Y^{obs}, Y^{mis})p(Y^{mis}|Y^{obs})dY^{mis} \\ &= \int I(Q = q(Y^{obs}, Y^{mis}))p(Y^{mis}|Y^{obs})dY^{mis}, \end{aligned} \quad (1)$$

where $I()$ is an indicator function taking the value 1 when the expression in parentheses is true and zero otherwise. A Monte-Carlo approximation to the posterior distribution of (1) is obtained by repeatedly drawing from $p(Y^{mis}|Y^{obs})$ and for each generated set of responses for the non-sampled individuals, $Y^{mis,*}$, computing $q(Y^{obs}, Y^{mis,*})$. Posterior summaries can then be approximated by the corresponding quantities in the Monte-Carlo sample. The Bayesian bootstrap (Rubin, 1981) which in the finite population case leads to a Polya urn sampling scheme for generating draws from $p(Y^{mis}|Y^{obs})$ provides a nonparametric approach to generating finite population Bayesian inferences (Lo, 1988; Meeden & Vardeman, 1991). Alternatively, under a parametric model indexed by θ , the predictive distribution of the unobserved responses is obtained via

$$p(Y^{mis}|Y^{obs}) = \int p(Y^{mis}|\theta, Y^{obs})p(\theta|Y^{obs})d\theta$$

and under standard modelling set-ups involving conditional independence of individual responses, given θ , this becomes

$$p(Y^{mis}|Y^{obs}) = \int p(Y^{mis}|\theta)p(\theta|Y^{obs})d\theta$$

where

$$p(\theta|Y^{obs}) \propto p(\theta) \prod p(Y_i^{obs}|\theta)$$

and Y_i^{obs} denotes the i th row of the observed data matrix, Y^{obs} . Under a parametric model, simulation of $p(Y^{mis}|Y^{obs})$ precedes in two steps: first draw θ^* from the posterior distribution for the parameters, then draw $Y^{mis,*}$ from the conditional predictive distribution $p(Y^{mis}|\theta = \theta^*)$. Under standard modelling assumptions the $N - n$, rows of each generated $Y^{mis,*}$ are obtained as independent draws from the common distribution $p(Y|\theta = \theta^*)$. In some situations, there may be covariate information available for all individuals in the population, in which case simulation of unobserved responses involves draws from $p(Y_i|\theta, X_i)$, for $i = n + 1, \dots, N$ and covariates X . In order to simplify notation this potential dependence on covariates is suppressed.

3.2 Bayesian predictive inference when the observed data cannot be released.

Now suppose (i) the data collector will not release the observed data, Y^{obs} , so that an analyst external to the data collector cannot compute the posterior distribution for Q and (ii) $N \gg n$, i.e. the sample size is a small fraction of the population size. Under condition (ii) and in view of the random sampling of the observed study sample it is reasonable to assume that the population quantity Q is closely approximated by the corresponding quantity computed from the responses for the non-sampled population and "filled-in" or imputed responses for the members of the observed study sample. That is, $q(Y^{pop}) = q(Y^{obs}, Y^{mis}) \approx q(Y^{imp}, Y^{mis})$ where the Y^{imp} are the imputed values which replace the observed data (Y^{obs}). Consequently, the posterior distribution for Q can be approximated via

$$p(Q|Y^{obs}) \approx \int I(Q = q(Y^{imp}, Y^{mis}))p(Y^{imp}, Y^{mis}|Y^{obs})dY^{mis} \quad (2)$$

and under standard modelling set-ups

$$\begin{aligned} p(Y^{imp}, Y^{mis}|Y^{obs}) &= \int p(Y^{imp}|\theta)p(Y^{mis}|\theta)p(\theta|Y^{obs})d\theta \\ &= \int \prod_{i=1}^n p(Y_i|\theta) \prod_{j=(n+1)}^N p(Y_j|\theta)p(\theta|Y^{obs})d\theta \\ &= \int \prod_{i=1}^N p(Y_i|\theta)p(\theta|Y^{obs})d\theta \end{aligned}$$

Note that the possibility of systematic differences between the observed study sample and the remainder of the population, induced by oversampling of some groups, can be accommodated in the above framework by explicit conditioning on the covariates used to determine selection probabilities. However, as noted above, we have left such conditioning implicit as a notational simplification.

In terms of Monte-Carlo approximation, the only change to the algorithm given after equation (1) above is that Q is computed from the generated values for the whole population rather than from the observed values for the study sample and generated values for the non-sampled population. This change makes it possible for the generation of the draws from $p(Y^{imp}, Y^{mis}|Y^{obs})$ and computation of the quantity of interest, Q , to be undertaken by separate parties, something which is not possible under (1) because the data collector will not release Y^{obs} . Under (2), the data collector could generate multiple draws from $p(Y^{imp}, Y^{mis}|Y^{obs})$ and an external analyst could compute desired quantities of interest for each generated dataset, thereby building up a Monte-Carlo approximation to (1) and hence to the posterior distribution of Q . This is the essence of

the multiple imputation approach to generating synthetic data-files: The data-collector uses the observed data to generate multiple draws from the posterior predictive distribution of unobserved values and external analysts are then free to analyse the generated ‘synthetic’ datasets. That is, synthetic datasets are generated as draws from a posterior predictive distribution and analysis of the repeated draws yields an approximation to the posterior distribution of the quantity of interest.

A generalisation of equation (2) is to base inference on

$$p(Q|Y^{obs}) \approx \int p(Q|Y^{imp}, Y^{mis})p(Y^{imp}, Y^{mis}|Y^{obs})dY^{mis} \quad (3)$$

This permits the uncertainty in approximating $q(Y^{pop})$ by $q(Y^{imp}, Y^{mis})$ to be accounted for. For example $p(Q|Y^{imp}, Y^{mis})$ might be approximated by a Normal distribution centred on $\hat{Q} = q(Y^{imp}, Y^{mis})$. This allowance for uncertainty may be relevant in cases where the sampling fraction is not negligible and the uncertainty arising from imputing responses for the observed study sample is appreciable.

The formulation in (3) is also the appropriate approximate posterior for situations where the analyst would compute a posterior distribution for Q (or, if adopting a frequentist approach, quote standard errors and confidence intervals) even if presented with the full population data Y^{pop} . As noted above, such situations arise when the objects of interest are parameters of statistical models and in some fields, such as epidemiology, modelling of data from complete populations such as a full birth cohort is relatively common. In these circumstance (3) represents an approximation to $p(Q|Y^{obs}) = \int p(Q|Y^{obs}, Y^{mis})p(Y^{mis}|Y^{obs})dY^{mis}$ and from the latter expression for $p(Q|Y^{obs})$ it can be seen that the approximation of (3) reflects the additional uncertainty due to imputing responses for the observed study sample.

While Monte-Carlo based inferences typically require thousands of samples to be drawn in order to provide accurate approximations, MI approaches exploit large-sample approximations to reduce the number of Monte-Carlo samples which must be generated by the data-collector and managed by an external analyst. For example in (3) a normal approximation would typically be adopted for $p(Q|Y^{imp}, Y^{mis})$, with suitable transformation of the quantity of interest in cases of obvious non-normality.

3.3 Bayesian predictive inference and practical synthetic data files

Henceforth let $Y^{gen} = (Y^{imp}, Y^{mis})$, for the responses of “generated population” comprising the imputed responses for the observed sample and actual responses for the non-sampled population.

When the population is large it may be impractical for data collectors to produce,

and external users to manage, multiple draws from $p(Y^{gen}|Y^{obs})$. Rubin and colleagues (Rubin, 1993; Reiter, 2002, Raghunathan et al, 2003) propose sub-sampling the draws from $p(Y^{gen}|Y^{obs})$ and releasing the sub-samples. Thus the proposal is to release random (usually, but not necessarily, simple random) samples from each draw from $p(Y^{gen}|Y^{obs})$. This is a pragmatic response to the practical difficulties of handling large datasets and it does not appear to be a necessary part of the MI approach to constructing synthetic data. If the population is not large then it may be feasible to release each complete draw from $p(Y^{gen}|Y^{obs})$ to users.

In order to relate the sub-sampling proposal to the Bayesian inference framework let Y^{sub} denote a random sample from Y^{gen} and note that although $p(Q|Y^{gen}, Y^{sub}) = p(Q|Y^{gen})$ it is nevertheless formally correct (though redundant) to write

$$p(Q|Y^{gen}) = \int p(Q|Y^{gen}, Y^{sub})p(Y^{sub}|Y^{gen})dY^{sub}$$

and therefore from (3)

$$p(Q|Y^{obs}) \approx \int \int p(Q|Y^{gen}, Y^{sub})p(Y^{sub}|Y^{gen})p(Y^{gen}|Y^{obs})dY^{sub}dY^{gen} \quad (4)$$

The proposal to release samples from the draws from $p(Y^{gen}|Y^{obs})$ amounts to replacing $p(Q|Y^{gen}, Y^{sub})$ by $(p(Q|Y^{sub}))$ in (4) and basing inference on

$$p^*(Q, Y^{obs}) = \int \int p(Q|Y^{sub})p(Y^{sub}|Y^{mis})p(Y^{mis}|Y^{obs})dY^{sub}dY^{mis}$$

Clearly, unless the sub-sampling fraction is high, $p(Q|Y^{sub})$ need not approximate $p(Q|Y^{gen}, Y^{sub}) = p(Q|Y^{gen})$ particularly closely. Since it conditions on a smaller size dataset, $p(Q|Y^{sub})$ will usually be more diffuse than $p(Q|Y^{gen})$. The implication of this is that inferences based on $p^*(Q, Y^{obs})$ will be less precise than those based on $p(Q|Y^{obs})$. This is simply the price that must be paid for working with sub-samples rather than full draws from $p(Y^{gen}|Y^{obs})$ and is analogous to the familiar issue of sampling variability. Just as a conventional analyst must (usually) work with Y^{obs} rather than Y^{pop} , an analyst of synthetic data must (usually) work with multiple simulations of Y^{sub} rather than multiple simulations of Y^{gen} .

As noted by Reiter (2005), in order to generate simulations of Y^{sub} it is not necessary to physically draw from $p(Y^{gen}|Y^{obs})$ and then draw a sample from the simulated Y^{gen} , because the combined effect of so-doing is identical to drawing from $p(Y^{sub}|Y^{obs})$. Under a standard parametric modelling set-up, draws from $p(Y^{sub}|Y^{obs})$ can be obtained by first drawing from the posterior distribution of model parameters, $p(\theta|Y^{obs})$ and then drawing from $p(Y^{sub}|\theta) = \prod_{i=1}^{n_{sub}} p(Y_i^{sub}|\theta)$. That is for each generated value of the model parameters a draw from $p(Y^{sub}|\theta)$ is obtained by independently sampling from

the $p(Y_i^{sub}|\theta)$, for $i = 1, \dots, n_{sub}$, where n_{sub} denotes the desired synthetic sample size. The generated draws from $p(Y^{sub}|Y^{obs})$ so obtained constitute the multiply imputed synthetic datasets released to analysts.

3.4 Inference given the synthetic data files

3.4.1 The analyst's inference problem

Although the preceding development provides some justification and motivation for the multiple imputation based approach to creating synthetic datasets it does not directly provide guidance as to how an analyst should proceed given multiple synthetic datasets, generated as draws from $p(Y^{sub}|Y^{obs})$. An analyst external to the data collector sees only the multiple synthetic samples released by the data collector and it is the analyst's task to quantify the information about Q contained in the synthetic files. A Bayesian analyst's task is to compute $p(Q|Y_{syn}^{sub})$, where Y_{syn}^{sub} denotes the collection of synthetic samples, $Y_1^{sub}, Y_2^{sub}, \dots, Y_M^{sub}$ obtained as independent draws from $p(Y^{sub}|Y^{obs})$. Given Y_{syn}^{sub} a non-Bayesian analyst may be satisfied with a method of computing point and interval estimates for Q , with good frequentist operating characteristics, where these operating characteristics are assessed over repeated sampling of Y^{obs} from Y^{pop} .

3.4.2 Approximate Bayesian inference

Raghunathan et al (2003) note that the posterior distribution for Q , conditional on the synthetic data, Y^{syn} can be decomposed as follows

$$p(Q|Y_{syn}^{sub}) = \int \int p(Q|Y_{syn}^{sub}, Y_{syn}^{gen}, Y^{obs})p(Y^{obs}|Y_{syn}^{sub}, Y_{syn}^{gen})p(Y_{syn}^{gen}|Y_{syn}^{sub})dY^{obs} dY_{syn}^{gen}$$

where Y_{syn}^{gen} denotes the collection of synthetic generated populations from which the the released synthetic samples are conceptually drawn. The synthetic populations are draws from $p(Y_{syn}^{gen}|Y^{obs})$. Because Y_{syn}^{sub} and Y_{syn}^{gen} are stochastic functions of the observed data, Y^{obs} , conditioning on them after conditioning on Y^{obs} , does not yield additional information and hence $p(Q|Y_{syn}^{sub}, Y_{syn}^{gen}, Y^{obs}) = p(Q|Y^{obs})$. Similarly $p(Y^{obs}|Y_{syn}^{sub}, Y_{syn}^{gen}) = p(Y^{obs}|Y_{syn}^{gen})$ and therefore

$$p(Q|Y_{syn}^{sub}) = \int \int p(Q|Y^{obs})p(Y^{obs}|Y_{syn}^{gen})p(Y_{syn}^{gen}|Y_{syn}^{sub})dY^{obs} dY_{syn}^{gen} \quad (5)$$

$$= \int p(Q|Y_{syn}^{gen})p(Y_{syn}^{gen}|Y_{syn}^{sub})dY_{syn}^{gen} \quad (6)$$

(Raghunathan et al 2003). Equation (5) illustrates the three sources of uncertainty which impact on inference based on synthetic data. The first component is posterior uncertainty given the observed data, $p(Q|Y^{obs})$. However because the observed data

is not available to analysts, it is, effectively, an unknown about which it is necessary to make inferences based on synthetic data. This additional source of uncertainty is represented in equation (5) by integration over the conditional distribution of the observed data, given the synthetic populations, $p(Y^{obs}|Y_{syn}^{gen})$. Finally, because it is not feasible to release the full synthetic populations, the additional uncertainty due to the release of samples from the synthetic populations is accounted for by the integration over $p(Y_{syn}^{gen}|Y_{syn}^{sub})$.

Using large sample normal approximations, Raghunathan et al (2003) derive the following approximation to the conditional mean and variance for Q , given Y^{syn} : Firstly, assume that given access to the original data, Y^{obs} , an analyst would summarise their information about Q , via a point estimate $q(Y^{obs})$ and a measure of uncertainty $v(Y^{obs})$. For example, a Bayesian may report the posterior mean and variance of Q , in which case both $q(Y^{obs})$ and $v(Y^{obs})$ may reflect prior information as well as the information in the observed data. In many large sample situations a normal distribution based on the posterior mean and variance will provide a reasonable approximation to the posterior for Q . For the m th synthetic sample, Y_m^{syn} , denote the corresponding point estimate and uncertainty measures by $q_m = q(Y_m^{sub})$ and $v_m = v(Y_m^{sub})$. Assuming M synthetic samples, Raghunathan et al (2003) derive a normal approximation to the posterior distribution $p(Q|Y_{syn}^{sub})$ with mean

$$\bar{q}_M = \frac{1}{M} \sum_m q_m \quad (7)$$

and variance

$$T_M = (1 + \frac{1}{M})b_M - \bar{v}_M \quad (8)$$

where $b_M = (M - 1)^{-1} \sum_m (q_m - \bar{q}_M)^2$ and $\bar{v}_M = \frac{1}{M} \sum_m v_m$. When the number of imputations is small it may be preferable to base inference on a t-distribution with mean and variance parameters as given above and degrees of freedom $df_M = (M - 1)(1 - r_M^{-1})^2$ where $r_M = (1 + M^{-1})b_M/\bar{v}_M$, (Reiter, 2002). The variance approximation given in (8) is not guaranteed to yield a positive variance, though this appears not to be an issue provided the size of synthetic datasets and the number of imputations is large (Reiter, 2002, 2005).

The variance formula (8) differs from the corresponding variance approximation for MI in conventional missing data problems because the average within imputation variance is subtracted rather than added to the between imputation variance. This difference reflects the sub-sampling of imputed populations in the creation of synthetic datasets. Intuitively, the subtraction of the average within imputation variance can be viewed as a correction to account for the fact that the size of the synthetic sub-samples can be determined by the imputer. Larger synthetic samples will lead to smaller between

imputation variance as well as smaller within imputation variance. Without some correction factor, large synthetic samples would over-state the inferential precision which should reflect the size of Y^{obs} , not the size of the synthetic samples which are generated from Y^{obs} . The subtraction of the average within imputation variance in (8) ensures that a synthetic sample size related decline in between imputation variance is balanced by a reduction in the magnitude of the negative contribution of the within imputation variance to the approximate posterior variance. The variance formula (8) was derived by Raghunathan et al (2003) as a simple approximation to the posterior variance conditional on the synthetic datasets, and these authors also suggest strategies for deriving more complex and accurate approximations to $p(Q|Y_{syn}^{sub})$. However the simple formula given in (8) has performed well in several simulation studies (Raghunathan et al, 2003; Reiter, 2002, 2005).

3.4.3 Approximate frequentist inference

Although the summary measures \bar{q}_M and T_M were derived via Bayesian arguments, they can also serve as the basis for frequentist inference: In a frequentist analysis the imputation specific point estimates and uncertainty measures would be standard estimates, such as maximum likelihood estimates and inverse information based variance estimates. A frequentist analyst can treat T_M as the estimated sampling variance of \bar{q}_M and base interval estimates on a normal or t-distribution centred on \bar{q}_M , with variance parameter, T_M . Raghunathan et al (2003) showed that, under certain conditions, such frequentist inference procedures are frequency valid in the sense that \bar{q}_M is an unbiased estimate of Q , T_M is an unbiased estimate of the sampling variance of \bar{q}_M and confidence intervals constructed from \bar{q}_M and T_M cover Q at the asserted level of confidence. The conditions under which these frequentist properties hold include unbiasedness of the point estimator for Q which would be computed given access to the observed data, $q(Y^{obs})$ and unbiasedness of the sampling variance estimator for $q(Y^{obs})$, again assuming access to the observed dataset. While these conditions could be expected to hold in many standard estimation problems, at least for large sample sizes, a further condition will not always hold as it relates to the propriety of the imputation procedure and is therefore a property of the imputation methodology.

The notion of proper imputation is, in general complex and subtle (Rubin, 1996; Meng 1994). In the current context it turns on the question of whether the value of Q which would be computed from each synthetic population (not sample), $q(Y_m^{gen})$ unbiasedly estimates the point estimate obtained from the observed data (Raghunathan 2003), that is whether $E(q(Y^{gen})|Y^{obs}) = q(Y^{obs})$. Because the synthetic populations will usually be constructed under some modelling assumptions, there is the potential for a given imputation procedure to be improper for a particular Q . This will occur if assumptions underlying construction of the synthetic datasets do not adequately

reflect features of the data which materially affect the value of Q . For example, suppose Q is a measure of association for two variables Y_1 and Y_2 , and that these variables are positively associated in the population and in the observed data so that $q(Y^{obs})$ takes some non-null value (e.g. odds ratio greater than one or a positive correlation coefficient). If the models underpinning the imputation methods assume Y_1 and Y_2 are independent, then apart from Monte Carlo sampling variation, Y_1 and Y_2 will tend not be associated in the synthetic populations and the association measure will not have expectation $q(Y^{obs})$. On the other hand, the very same imputation procedures may be proper, and inference procedures based on (7) and (8) may be frequency valid, for features of the marginal distributions of Y_1 and Y_2 . The propriety and frequency validity of imputation procedures are specific to each quantity of interest.

Based on results from the application of MI to conventional missing data problems it seems reasonable to conjecture that some synthetic data imputation methods may be technically improper but not harmful to the frequency validity of estimation procedures. (Rubin, 1996, Meng 1994). For example, suppose the imputer has strong prior information that Y_1 and Y_2 are associated and reflects this prior information in the prior distributions adopted for the models underpinning the imputations. Suppose further that in the sample actually observed Y_1 and Y_2 are not associated, although this is due to sampling variability and Y_1 and Y_2 are, in fact, associated in the population. The imputed synthetic populations will reflect both the imputer's prior model and the observed data and so could be expected to exhibit association between Y_1 and Y_2 , although the strength of this association is likely to be less than assumed by the imputer's prior. If the analyst adopts an uninformative prior for the association between Y_1 and Y_2 , the point estimate they would obtain from the observed data would reflect the absence of association exhibited in the observed data. Under the same uninformative prior, estimates obtained from the synthetic populations would reflect the association in those populations and would therefore not unbiasedly estimate $q(Y^{obs})$ so that the imputation procedure is improper. Nevertheless, assuming the imputer's prior assumption is, in fact correct, and Y_1 and Y_2 are associated in the observable population data the frequency operating characteristics of the analysts inference procedures, given the synthetic data, will be superior to inferences based on proper imputation procedures, designed so that synthetic population association measures unbiasedly estimate the (incorrect) null value observed in the observed study sample. For example, the bias of analyses based on the synthetic data should be less than the bias for the same analyses of the observed data.

3.5 A synthetic data framework for census data.

Sometimes there may be interest in constructing a synthetic version of a dataset which represents a complete survey of a population rather than a sample survey. For example, a national statistical office may be interested in releasing synthetic versions of a population census dataset. Because the preceding development emphasised the logic of sample to population inference, the construction of synthetic census datasets lies strictly outside the conceptual framework developed thus far. However, there appear to be at least two ways in which an MI approach to constructing synthetic census datasets may be motivated and justified. Firstly, an observed population census dataset could be regarded as having been obtained via sampling from some much larger hypothetical "super-population." Though somewhat contrived, this idea seems implicit in the work of modellers who fit models to census datasets and focus inferential attention on the model parameter. Parameters of statistical models are generally interpretable as long-run limits of functions of observables and such long-run limits are, in essence, another way of defining the notion of super-population. Under a super-population model the framework outlined in sections 3.1 to 3.4 can be applied.

A second approach to motivating the construction of synthetic versions of census datasets may be more appropriate when the objects of interest are observable features of a finite population, such as the proportion of the population who smoke. An observed census dataset for a finite population of size N , could be regarded as recording the responses of the first N individuals in an exchangeable sequence of individuals. Denote the unobserved data for the next N individuals in the sequence by Y^{new} . Under exchangeability it is reasonable to assume that for large N , $Q = q(Y^{obs}) \approx Q_1 = q(Y^{new})$, for any quantity of interest, Q . So inference for Q_1 is approximately also inference for Q and the imputer can generate and release synthetic datasets by drawing from $p(Y^{new}|Y^{obs})$. If the N future individuals are regarded as a defined finite population, then this proposal differs from that outlined in sections 3.1 to 3.4 because it does not involve sub-sampling of the draws from the posterior predictive distribution of the data for the unobserved population. The practical implication of this is that the formula for computing the approximate posterior variance from multiple synthetic datasets will differ from (8) because the fact that each synthetic dataset represents an imputation of a full population means that the within imputation variances will be zero and the variance approximation will reflect only between imputation variance. However, the future population size need not be fixed at the size of the observed population and is, in fact, entirely in the hands of the imputer. Consequently, it may be more justifiable to regard the generated synthetic future populations as samples from an infinite exchangeable sequence. Under this interpretation the variance formula (8) applies and it can be seen that the two approaches to justifying synthetic versions of census datasets lead to imputation and

analysis methods which are operationally identical. The primary difference between the two approaches is therefore the conceptual issue of whether the object of inferential interest is a super-population parameter (such as a model parameter) or a descriptive statistic for a finite population.

3.6 Constructing imputation models

As noted in section 3.1, computation of the posterior predictive distribution of the unobserved data usually requires some modelling assumptions. In the context of generating synthetic data files the development of plausible statistical models for the observed data is a crucial step. Imputations based on erroneous modelling assumptions will reflect those assumptions and are likely to lead to misleading inferences.

Given a model parameter, θ , we assume the rows of Y^{obs} are conditionally independent and identically distributed as $p(Y|\theta)$. The most general model for $p(Y|\theta)$ is the unrestricted multinomial defined over the possible combinations of variables occurring in the population. This leads to imputation procedures based on the Bayesian bootstrap (Rubin, 1981, Lo 1988). Because these procedures sample complete records from the observed data file, they are potentially problematic from a disclosure limitation viewpoint. At the other extreme from the Bayesian bootstrap are fully parametric procedures. When the dimension of Y is small it may be possible to specify a single multivariate model for Y but in many realistic situations it will be convenient to model Y via a sequence of conditional distributions. Suppose the components of Y can be divided into J groups of variables and that $\theta = (\theta_1, \theta_2, \dots, \theta_J)$, with

$$p(Y|\theta) = p(Y_1|\theta_1)p(Y_2|\theta_2, Y_1), \dots, p(Y_J|\theta_J, Y_1, \dots, Y_{J-1}) \quad (9)$$

If accompanied by an *a priori* assumption of mutual independence for the J components of θ this leads to *a posteriori* independence for the components of θ , which simplifies simulation. That is, under the partition (9), *a priori* independence for the J components of θ ensures

$$p(\theta|Y^{obs}) = \prod_j p(\theta_j|Y_{\cdot,1}^{obs}, \dots, Y_{\cdot,j}^{obs})$$

and draws from $p(\theta|Y^{obs})$ can be obtained by simulating independently from the component posterior distributions, $p(\theta_j|Y_{\cdot,1}^{obs}, \dots, Y_{\cdot,j}^{obs})$. Here, the notation $Y_{\cdot,j}^{obs}$ denotes the columns of the observed data matrix corresponding to the j th group of variables.

Specification of the sequence of conditional distributions will reflect convenience and prior knowledge. For example, knowledge of conditional independencies between the variables can simplify some of the component models and logical orderings between variables can be exploited in determining the order of the conditional models. Reiter (2005) gives an example involving the generation of synthetic data via a sequence of

seven univariate conditional regression models.

When a group of non-sensitive variables, posing little confidentiality risk, can be identified, the modelling task could be simplified by modelling these variables non-parametrically and building parametric models for the remaining variables conditionally on these low-risk variables. This leads to a semi-parametric approach in which part of each synthetic record is obtained via Bayesian bootstrapping from the observed dataset, with the remainder of each record generated, under a sequence of parametric models, via draws from the posterior predictive distributions of the relevant variables.

One modelling strategy which has not yet been evaluated in the context of synthetic data is to lessen the impact of modelling assumptions by employing hierarchical Bayesian (HB) models for at least some of the component models. Hierarchical generalised linear models (Albert 1988), embed the usual structural part of the GLM formulation, relating conditional means to covariates, within a probabilistic structure in which the usual GLM model serves as the prior mean for the conditional expectation of the response variable. This results in posterior estimates which are compromises between simple data-based estimates and those that would obtain under full prior commitment to a specified GLM. In the context of synthetic dataset creation, this should give some protection from a poor choice of GLM because in an HB framework the observed data can pull posterior estimates away from the specified GLM formulation. An example of the application of a hierarchical Bayesian Poisson GLM to the generation of synthetic data is given in section 4, below.

4 Example: The institutional care data

4.1 Background

In order to illustrate the multiple imputation based approach to constructing and analysing synthetic data we used a dataset concerned with social variation in the prevalence of utilisation of institutional care, defined as permanent residence in a rest home or hospital. There are only five variables on the dataset, age (in five year categories), sex, ethnicity, (Maori, Pacific, and non-Maori, non-Pacific) educational achievement (no qualifications, school qualifications, trade qualifications, undergraduate and post-graduate degree qualifications) and an indicator for permanent residence in a health care institution (yes/no). Because of the small number of variables and the fact that they are all categorical, this example dataset represents a special case of synthetic data creation. Nevertheless, it serves to illustrate many of the essential ideas of the multiple imputation approach to constructing synthetic data files. The fact that all variables are categorical means that the dataset, comprising in excess of 1.95 million records can be represented as a five-way table comprising 480 cells and this simplifies the construction

of an imputation model, as discussed below in sections 4.2 and 4.3. However, we emphasise that, in this example, there is no intrinsic difference between the unit record file comprising 1.95 million records and the tabular dataset in which each record is allocated to one of the 480 possible combinations of variable values. The information content of both data representations is necessarily identical and it is a trivial matter to transform one representation to the other. We use the tabular representation henceforth because this simplifies modelling, however given imputed synthetic tables an imputer or analyst could easily produce the equivalent synthetic sets of unit record files.

The dataset used as the observed data in this application was based on an extract from the New Zealand Census-Mortality Study, database, which links mortality records to the 1996 population census. (Hill, Atkinson & Blakely, 2002). The data were originally extracted in connection with a study of social variation in health expectancy, and were restricted to ages 25 to 74 to match age-restrictions of other data sources used in that study. The data extracted comprised institutionalisation prevalences and denominator counts for each age-sex-ethnicity-educational achievement group. The data were originally accessed in the Statistics New Zealand datalab facility and, before extraction, were confidentialised by rounding of prevalences of institutional care and random rounding of denominator counts. That is, the dataset taken as the ‘observed data’ in this example was, in fact, a modified version of the original dataset. Approximate counts of people in institutional care were constructed by multiplying the rounded prevalences by the randomly rounded denominator counts, and rounding the result to the nearest integer. Counts of people not in institutional care were obtained by subtracting the approximate counts in institutional care from the corresponding randomly rounded denominators. This created a dataset comprising cell counts for a five way cross-classification of age, sex, ethnicity, educational achievement and institutional care utilisation. and it was this dataset which served as the observed data for the analyses reported below. We emphasise that because of the rounding and random rounding, cell counts in our ‘observed’ data do not reflect the actual counts in the census database.

In many ways the institutional care data is a challenging dataset from both confidentiality and analytical perspectives. The prevalence of institutional care in the 25 to 74 age-group is only 0.3% and combined with the uneven distribution of educational qualifications and ethnicity this leads to several cells with small counts. In particular the cross-classified dataset contains 53 zero cell counts and 28 cells with a count of one. The ethnic groups range in size from 1,664,481 for non-Maori, non- Pacific to 219,126 for Maori and 74,841 for Pacific people. The distribution of educational achievement also varied markedly by ethnicity. For example while 10.9% of the non-Maori, non-Pacific group had a tertiary qualification, only 3.1% of the Maori group and 2.7% of the Pacific group were tertiary qualified. Numbers in institutional care within some sub-groups, are low even after collapsing over the age variable. For example, only five

tertiary qualified Maori and four tertiary qualified Pacific people were recorded as in institutional care

4.2 Imputation models for categorical data: Some background.

4.2.1 Poisson log-linear models

Because of the small number of variables involved in this application and the fact that they are all restricted to take only a small number of possible values it is possible to model the joint distribution of the variables via a single log-linear model for the cell counts in the table formed by cross-classification of the variables. That is, in the notation established in section 3, this is a case where the joint distribution $p(Y|\theta)$ can be modelled via a single model rather than a sequence of conditional models.

For cell counts C_i , $i = 1 \dots K$, a conventional Poisson log-linear model can be specified via

$$C_i | \lambda_i \sim \text{Poisson}(\lambda_i) \quad (10)$$

$$\ln(\lambda_i) = X_i \beta \quad (11)$$

where X_i is a vector of variable values descriptive of the i^{th} cell and β is a vector of unknown model parameters, to be estimated from the data. We will use an underscore notation such as \underline{C} or $\underline{\lambda}$ to denote $K - fold$ vectors comprising the corresponding cell-specific variables for all K cells of the cross-classification. For the institutional care data, the covariate vector X_i includes age category, sex, ethnicity, educational achievement and an indicator for permanent residence in a health care institution, as well as product terms involving various combinations of these variables. Details of the model formulations are given below in section 4.3.1. Conditionally on the model parameter β , the Poisson log-linear model implies the expected cell counts are $E(C_i | \lambda_i) = \exp(X_i \beta)$, and given an estimate of β , $\hat{\beta}$, say estimated or fitted cell counts are given by $\hat{\lambda}_i = \exp(X_i \hat{\beta})$, which exactly reproduce the assumed log-linear model form. Moreover, any assumptions of independence or conditional independence between variables, represented by the absence of certain interaction terms (product terms) in the model specification will be exactly reproduced in the fitted values. Except for Monte Carlo sampling variation, the same holds for synthetic datasets generated via posterior predictive inference based on a log-linear model.

Under the Poisson log-linear model, synthetic datasets would be generated by first drawing from the posterior distribution of the model parameters and for each generated value, β^* , computing predicted cell means $\lambda_i^* = \exp(X_i \beta^*)$, for $i = 1, \dots, K$, and finally generating synthetic cell counts by independent draws from the Poisson distributions, $C_i^* | \lambda_i^* \sim \text{Poisson}(\lambda_i^*)$, for $i = 1 \dots K$. Omitted interaction terms in the fitted

model, correspond to setting certain elements of the parameter vector to zero and in the Monte Carlo simulation this means that those elements are zero for every draw from the posterior. Consequently each set of expected cell counts $(\lambda_i^*, i = 1, \dots, K)$ generated in the Monte Carlo simulation will exactly reproduce both the assumed log-linear model form and assumptions regarding absence of interaction. The final step of the simulation induces Poisson variation around the expected cell counts, so individual synthetic datasets may deviate from the assumptions of the assumed log-linear model although averages taken over the synthetic datasets will reflect the assumptions of the fitted model.

4.2.2 Hierarchical Poisson log-linear models

Basic ideas. The impact of the log-linear modelling assumptions on fitted values and therefore on the structure of synthetic datasets can be lessened by embedding the conventional log-linear model in a hierarchical structure in which the log-linear formulation specifies the means for prior distributions for the expected cell counts, λ_i , rather than directly determining these expectations. A hierarchical Bayes version, of the Poisson log-linear model can be written

$$C_i | \lambda_i \sim \text{Poisson}(\lambda_i)$$

$$\lambda_i | X_i, \beta, \xi \sim \text{Gamma}(\xi, \xi / \mu_i) \quad (12)$$

$$\ln(\mu_i) = X_i \beta \quad (13)$$

$$(\beta, \xi) \sim \pi \quad (14)$$

where π denotes the prior distribution of the parameters of the Gamma-prior for the cell means. The parameters of the prior model for the expected cell counts (β, ξ) are often referred to as hyper-parameters, to distinguish them from the expected cell counts themselves (i.e. the λ_i). Similarly, the prior for the hyper-parameters (π) is referred to as the hyper-prior to distinguish it from the Gamma prior for the expected cell counts. The Gamma prior is chosen because it is conjugate to the Poisson distribution and this simplifies computation. Under the parameterisation adopted in (12) the Gamma prior at the second level of the hierarchical model (12) has mean μ_i which is related to the covariates (descriptions of cell characteristics) via the log-linear model of (13). This illustrates the point made above that in moving to a hierarchical model, the requirement to specify a functional form for the dependence of expected cell counts is moved up a level, and now only determines the prior mean for the expected cell counts, rather than directly determining the form of the expected cell counts. A corollary of this is that the observed data may shift the posterior for the λ_i away from the specified prior functional form. In fact, fitted values under the hierarchical Bayesian log-linear model

are, approximately, compromises between the observed cell counts and the fitted values under the prior GLM model, as discussed below (see “shrinkage estimation”).

The variance of the Gamma prior for the expected cell counts (12) is μ_i^2/ξ . The parameter ξ can be viewed as an index of the degree of confidence in the assumed model for the expected cell counts. In particular, as ξ becomes large the prior variance around μ_i becomes small and in the limit, as ξ approaches infinity, the hierarchical model formulation reduces to the standard GLM formulation, given by equations (10) and (11). It should be noted, however, that in a fully Bayesian analysis ξ is not specified *a priori*, but as with the regression parameters, β , it is assigned a prior distribution. It follows that the observed data are allowed to determine the adequacy of the posited prior model, particularly if, as is often the case in practice, a vague prior for ξ is specified.

Shrinkage estimation Some insight into the implications of hierarchical Bayesian modelling for the generation of synthetic datasets can be obtained by considering the conditional posterior distributions for the expected cell counts, given the hyperparameters (β, ξ) , even though full Bayesian inference ultimately requires accounting for the uncertainty concerning the hyperparameters by integrating over their posterior distribution.

Given the hierarchical structure of the Gamma-Poisson model given by equations (3) and (4), the conditional posterior distribution for the expected cell count is also Gamma distributed, i.e.

$$\xi_i | \underline{C}, \beta, \xi \sim \text{Gamma}(\xi + C_i, \xi/\mu_i + 1) \quad (15)$$

with mean

$$E(\lambda_i | \underline{C}, \beta, \xi) = B_i \mu_i + (1 - B_i) C_i \quad (16)$$

where $B_i = \xi/(\xi + \mu_i)$ (Christiansen & Morris, 1997). That is, the conditional posterior mean for the i th expected cell count (λ_i) is a weighted average of the prior mean (μ_i) and the cell count C_i . Moreover, as ξ becomes large the weight on the prior mean approaches one, whereas as ξ approaches zero the weight on the prior mean approaches zero and the weight on the cell count approaches one.

The fact that the conditional posterior mean for each expected cell count is a weighted average of the observed cell count and the prior mean under the posited log-linear model (μ_i) has the interesting implication of ensuring that the conditional posterior means always lie closer to the observed cell counts than do the prior means. If the conditional posterior means for the expected cell-counts are regarded as fitted values under the hierarchical model, it can be seen that assumptions of smoothness or absence of interaction built into the prior model will not necessarily be exactly repli-

cated in the fitted values, in contrast to the situation with conventional GLMs. In terms of generating synthetic datasets the implications of this are that averages taken over multiply imputed synthetic datasets generated under a hierarchical model will not usually exactly reflect the assumptions of the prior model. This will be particularly true in regions of the data which are information-rich. On the other hand, when data are sparse, the prior model can be expected to have considerable influence on posterior estimates, and hence the structure of synthetic datasets.

Specification of the hyper-prior In order to complete the specification of the hierarchical model the hyper-parameters need to be assigned a prior distribution, the "hyper-prior." Uninformative choices for the hyper-prior are commonly employed in applications and we adopted this convention in developing imputation models for the institutional care data. In particular, we assumed a uniform prior for the regression hyper-parameter, β , and a uniform shrinkage prior for the precision hyper-parameter, ξ . The latter prior is derived via transformation of a uniform prior for a shrinkage parameter (Christiansen & Morris, 1997; Daniels, 1999). If we consider, for the moment, a single cell, then assuming the shrinkage parameter $B_i = \xi/(\xi + \mu_i)$ has a Uniform prior on $[0,1]$ implies $p(\xi) \propto \mu_i/(\mu_i + \xi)^2$. Because this prior depends on the specific cell in question, via μ_i , the prior for ξ is actually derived by replacing μ_i by a constant, z_0 , say (Christiansen and Morris 1997). This yields $p(\xi) \propto z_0/(z_0 + \xi)^2$. The parameter z_0 can be thought of as the expected cell count in some reference cell, labeled 'cell 0'. It is also the prior median of ξ and larger values of z_0 encourage more shrinkage than smaller values (Christiansen and Morris 1997).

Various strategies for setting z_0 can be contemplated (Daniels 1999), however we have found it useful to focus on the corresponding shrinkage parameter and to fix z_0 at the expected cell count for which a uniform prior on the shrinkage would be reasonable. At this expected cell count we are ambivalent about the relative weight which should be given to the prior model and, in fact our expected prior shrinkage (conditional on the hyper-parameters) is 0.5. It follows that at expected cell counts larger than z_0 we expect the model based estimates to be given less weight than the observed counts, and the model to be given more weight than the observed counts at smaller expected cell counts. Sensitivity to the choice of z_0 can be addressed via sensitivity analysis. However, in practice, we have found posterior inferences to be insensitive to the specific choice of z_0 .

4.3 Generation of synthetic datasets

4.3.1 Log-linear model specifications for imputing the institutional care

data.

Complex imputation model Models to be used as the basis for synthetic data need to fit observed data well, while permitting some smoothing, since the latter implies some variation between and observed and fitted values and hence variation between observed and average synthetic values, and this increases confidentiality protection. Models used to multiply impute synthetic data will typically be more complex than models that would be fitted for analytical purposes but must omit some complex interactions to avoid saturating the observed data. The principal log-linear structure adopted for modelling the institutional care data included all possible interactions between age, sex, ethnicity and educational achievement and three way interactions involving all two-way combinations of these variables and the institutional care indicator. Age was treated as a continuous variable in these model specifications and was represented via a quadratic spline with a single knot at age 50 (Greenland, 1995), except that, as a simplification, only the linear part of the age variable was included in three-way interaction terms involving age and institutionalisation. In addition, for the three way interaction between ethnicity, education and institutionalisation, the trade and tertiary qualifications categories were combined into a single category, for Pacific people only. Numerical problems were encountered in trying to fit the full three-way ethnicity-education - institutionalisation interaction. This presumably reflects the small number of Pacific people with degree level qualifications in 1996 and the low prevalence of institutional care.

The structure just described served as the model for the prior means for the expected cell counts, (i.e. for the μ_i) in the hierarchical model and as the model for the expected cell counts, (i.e. for the λ_i) in the conventional log-linear model (11). The model is close to a log-linear model representation of a logistic model for institutionalisation including age, sex, ethnicity and education and two way interactions terms between these variables as predictors (Agresti, 1990 pp 177-178). However, because age was treated as a continuous variable rather than as a 10-category categorical variable the age by sex by ethnicity by education sub-table is not saturated under the log-linear model specification adopted here.

Simple imputation model In order to explore the impact of the choice of imputation model we also generated synthetic datasets under a simpler imputation model which differed from the model described above by including only two-way interactions involving institutional care. This approximates a logistic regression model for institutionalisation which includes age, sex, ethnicity, and education as predictors but excludes interaction terms involving these variable. This model could be described as a "main effects logistic regression model". Note however as with the complex imputation model the age by sex by ethnicity by education sub-table is not saturated under the simple

imputation model.

4.3.2 Model-fitting

We used an R implementation of Christiansen and Morris's S-plus function PRIMM (Christiansen & Morris, 1997) in order to fit hierarchical Bayesian Poisson log-linear models to the institutional care data. Conventional non-hierarchical versions of the models were fitted using the R function, `glm`

4.3.3 Simulation algorithm.

The generation of synthetic datasets followed the logic of Bayesian predictive inference, by generating draws from the posterior predictive distribution of new data given the observed data. Because, in this example, the data can be represented as a multiway table, synthetic datasets can be generated by drawing new or synthetic cell counts, from $p(\underline{C}^{sub}|\underline{C}^{obs})$ where \underline{C}^{sub} and \underline{C}^{obs} denote, respectively, vectors of new (or synthetic) and observed cell counts corresponding to the $K = 480$ cells of the cross-classification. Under the hierarchical Bayesian model

$$p(\underline{C}^{sub}|\underline{C}^{obs}) = \int \int \int \prod_i p(C_i^{sub}|\lambda_i)p(\lambda_i|C_i^{obs}, \beta, \xi)p(\beta, \xi|\underline{C}^{obs})d\lambda, d\beta, d\xi$$

and the algorithm for generating the synthetic datasets simply approximates this integral by drawing from the relevant distributions for each component of the integrand, starting with the right-most component and moving to the left-most component. That is, under the hierarchical Bayesian model, synthetic datasets are generated via repetition of the following steps.

1. draw (β^*, ξ^*) from $p(\beta, \xi|\underline{C}^{obs})$
2. For $i = 1 \dots K$, independently draw λ_i^* from $p(\lambda_i|C_i^{obs}, \beta = \beta^*, \xi = \xi^*)$
3. For $i = 1 \dots K$, independently draw $C_i^{sub,*}$ from $p(C_i^{sub}|\lambda_i = \lambda_i^*)$

Each cycle through steps (1) to (3) generates a synthetic dataset. In step 1 a multivariate normal approximation was adopted for $p(\beta, \ln(\xi)|\underline{C}^{obs})$, (Christiansen & Morris, 1987). The mean and covariance matrix on which to base this approximation can be extracted from the PRIMM output. In step two, the expected cell counts are drawn from the independent Gamma distributions given in (15). Under the HB model, step three involves drawing independently from Poisson distributions, with expected values given by the λ_i . For the analyses reported below we generated 100 synthetic datasets. We constrained the size of each synthetic dataset (i.e. the sum of cell counts) to equal the size of the observed dataset by replacing step 3 with a draw from a multinomial

distribution, with sample size set to the size of the observed data and cell probabilities given by $\lambda_i^*/\sum_i \lambda_i^*$, for $i = 1 \dots K$.

Under a non-hierarchical model, step one of the above algorithm involves drawing from the posterior distribution of the log-linear model parameters, conveniently approximated as multivariate normal centred on the maximum likelihood estimate with variance given by the usual inverse information based variance estimate. In addition, for a non-hierarchical model, step two of the above algorithm is replaced by a deterministic step: $\lambda_i^* = \exp(X_i\beta^*)$.

4.4 Analysis of synthetic datasets

4.4.1 Impact of synthetic data on inference

We compared model fit diagnostics and parameter estimates obtained from fitting logistic models predicting institutional care prevalence to the observed data with those obtained from fitting the same models to synthetic data, generated under both conventional (GLM) and HB log-linear imputation models. For the synthetic data analyses, point estimates and standard errors were obtained under the combining rules (7) and (8) given in section 3.4. As a point of comparison, we also fitted the logistic models to a randomly rounded version of the observed data. Random rounding is the current Statistics New Zealand standard for confidentialising datasets which, as in this example, can be represented in tabular form. The current Statistics New Zealand random rounding protocol leaves unchanged cell counts which are multiples of 3, but randomly rounds other counts, either to the nearest multiple of 3 less than the observed cell count or to the nearest multiple of 3 greater than the observed cell count, with probability inversely proportional to the distance between the observed cell count and the proximate multiples of 3.

We fitted two logistic regression models to the observed, randomly rounded and synthetic datasets. The first model (the "main effects" model) included age, sex, ethnicity and educational achievement as predictors but omitted interactions between these variables. Educational achievement was represented by three dummy variables for no qualifications, trades qualifications and university level qualifications, leaving secondary school qualifications as the reference category. Ethnicity was represented via two dummy variables for the Maori and Pacific Island groups, leaving the non-Maori, non-Pacific group as the reference category. The second model (the "ethnicity interaction model") was more complex and included ethnicity by age, ethnicity by sex, and ethnicity by educational achievement interaction terms. For the latter, the trades and university qualifications categories were collapsed into a single post-secondary category. An explanation of the variable abbreviations used throughout the analysis is given in Table 1.

Table 1. Variables used in the logistic regression analyses

Variable	Definition
age	age - coded as a continuous variable
sex	indicator for male sex
M	indicator for Maori ethnicity
PI	indicator for Pacific ethnicity
nq	indicator for no qualifications
trade	indicator for trade qualifications
tert	indicator for degree level qualifications
am	age \times M -age by Maori ethnicity interaction term
api	age \times PI -age by Pacific ethnicity interaction term
msex	M \times sex - Maori ethnicity by sex interaction term
pisex	PI \times sex - Pacific ethnicity by sex interaction term
manq	M \times nq - Maori ethnicity by no qualifications interaction
mapsec	M \times (trade + tert) Maori ethnicity by post-secondary interaction
pinq	PI \times nq - Pacific ethnicity by no qualifications interaction
pipsec	PI \times (trade + tert) Pacific ethnicity by post-secondary interaction

Deviance statistics (-2 times the log-likelihood) for the main effects and ethnicity interaction model, fitted to observed, randomly rounded and synthetic data, created under the simple and complex GLM and HB models, are reported in Table 2. Also shown are the reductions in the Akaike Information Criterion (AIC), defined as minus two times the log-likelihood plus twice the number of parameters in the model. The AIC adds a penalty for model complexity to the deviance statistics which decline as additional parameters are added to a model. A smaller AIC indicates a preferable model. For the synthetic data analyses the deviance statistics were obtained as the average over the deviance statistics for each of the 100 synthetic datasets generated.

The deviance statistics obtained from modelling the randomly rounded data were greater than those obtained from the observed data, whereas the deviance statistics for the multiply imputed synthetic datasets were smaller than the corresponding statistics for the observed data. This presumably reflects the additional irregularity induced by random rounding and the smoothing induced by the imputation models which underpin the synthetic datasets. The deviance statistics obtained from modelling the HB based synthetic datasets were closer to the corresponding statistics for the observed data than were the deviances obtained from modelling the GLM based synthetic datasets. This is consistent with the notion that HB models provide a compromise between data and an assumed GLM.

The effect of the imputation model on deviance statistics was particularly strong for the simple GLM imputation model. For synthetic datasets generated under this imputation model, not only were deviance statistics substantially reduced compared to

the deviance for the observed data, but the comparison of the fit of the main effects and ethnicity interaction analysis models yielded a slight preference in favour of the main effects model. This stands in contrast to the situation for the observed data for which the comparison of analysis models suggests a strong preference for the ethnicity interaction model. The structure of the simple GLM imputation model was close to that implied by the main-effects logistic regression model and it is therefore not surprising that fitting the latter model to synthetic data generated under the simple GLM imputation model yields a good fit. The residual degrees of freedom for the main effects logistic regression model is 232 and for a model fitted to conventional non-synthetic data the residual deviance of 217 observed for the simple GLM based synthetic data gives a p-value of 0.752, which would usually be taken to indicate a more than satisfactory degree of congruence between data and model. (Distributional results for deviance statistics computed from multiply imputed synthetic data are not currently available). Given this evidence of good fit for the main effects logistic model, it is not surprising that fitting the ethnicity interaction model to the simple GLM based synthetic datasets fails to markedly improve on the fit of the main effects model.

In contrast to the situation for the synthetic data based on the simple GLM imputation model, deviance and AIC statistics for models fitted to synthetic datasets created under the HB analog of the simple GLM model provided clear support for the ethnicity interaction model. However the magnitude of this support was substantially less than indicated by the observed data.

For synthetic data created under both HB and GLM versions of the complex imputation models, comparison of deviance statistics obtained from fitting the main effects and ethnicity interaction models suggests a similar level of support for the ethnicity interaction model to that indicated by the analysis of the observed data. The results reported in Table 2 suggest that goodness of fit for models fitted to multiply imputed synthetic datasets created under HB imputation models are less sensitive to the assumptions of the imputation model concerning the form of associations between variables than is the case for the corresponding GLM imputation models.

Table 2. Deviance and AIC statistics for main effects and ethnicity interaction model fitted to observed data (Obs), randomly rounded data (RR), and synthetic data generated under simple and complex GLM (Syn-GLM) and HB models (Syn-HB)

data type	Deviance-1 ¹	Deviance-2 ²	Deviance Reduction	AIC Reduction
Observed	1866.7	1795.3	71.4	55.4
Randomly rounded	1955.8	1872.2	83.6	67.6
Syn-GLM simple	217.0	209.0	8.0	-8.0
Syn-GLM complex	1424.4	1336.5	87.9	71.9
Syn-HB simple	1527.8	1491.6	36.2	20.2
Syn-HB complex	1685.3	1597.5	87.7	71.7

¹Deviance for main effects logistic model (8 parameters)

²Deviance for ethnicity interaction model (16 parameters)

Parameter estimates and standard errors obtained from fitting the main effects and ethnicity interaction models to the observed, randomly rounded and synthetic data generated under the conventional ("Syn-GLM") and hierarchical ("Syn-HB") versions of the complex log-linear imputation model are shown in Tables 3 (main effects model) and 4 (ethnicity interaction model). The results reported in Table 3 indicate that fitting the main effects logistic model produces similar parameter estimates whether fitted to the observed data or to multiply imputed synthetic versions of the data, generated under either the conventional or hierarchical versions of the complex log-linear model. When applied to the randomly rounded data the main effects logistic regression model produced an estimate for the effect of Pacific Island ethnicity which differed by approximately one standard error from the estimates obtained from the other datasets. Other parameter estimates obtained from fitting the main effects logistic model to the randomly rounded data were similar to those obtained from the original and synthetic datasets.

Table 3. Comparison of logistic regression parameter estimates ($\hat{\beta}$) and standard errors (se) obtained from fitting a simple "main effects" model to observed (Obs), randomly rounded (RR) and synthetic data, imputed under complex non-hierarchical (Syn-GLM) and hierarchical (Syn-HB) log-linear models

	Obs		RR		Syn-GLM		Syn-HB	
	$\hat{\beta}$	se	$\hat{\beta}$	se	$\hat{\beta}$	se	$\hat{\beta}$	se
β_0	-6.18	0.04	-6.17	0.04	-6.18	0.04	-6.19	0.03
β_{age}	0.06	0.001	0.06	0.001	0.06	0.002	0.06	0.001
β_{sex}	-0.18	0.03	-0.19	0.03	-0.19	0.03	-0.18	0.03
β_M	-0.00	0.04	0.01	0.04	-0.00	0.04	0.01	0.04
β_{PI}	-0.30	0.08	-0.23	0.08	-0.30	0.09	-0.29	0.07
β_{nq}	0.93	0.04	0.92	0.04	0.93	0.05	0.93	0.02
β_{trade}	-0.15	0.05	-0.15	0.05	-0.14	0.06	-0.15	0.05
β_{tert}	-0.84	0.10	-0.84	0.10	-0.83	0.12	-0.82	0.11

While fitting the ethnicity interaction model to the original and multiply imputed synthetic datasets, (Table 4) produced more variation in parameter estimates than for the main effects logistic model, these differences were generally small, relative to the standard errors. The randomly rounded data again produced an estimate of the Pacific Island main effect parameter which differed markedly from those obtained from the observed and multiply imputed data. The point estimate for randomly rounded data differed by 0.25 from the estimate based on the observed data, with standard errors of 0.22 and 0.21 for the observed and randomly rounded data respectively. Parameter estimates for the Pacific Island - no qualifications interaction term also differed by almost one standard error between the original and randomly rounded data.

Table 4. Comparison of logistic regression parameter estimates ($\hat{\beta}$) and standard errors (se) obtained from fitting the “ethnicity interaction” model to observed (Obs), randomly rounded (RR) and synthetic data, imputed under complex non-hierarchical (Syn-GLM) and hierarchical (Syn-HB) log-linear models

	Obs		RR		Syn-GLM		Syn-HB	
	$\hat{\beta}$	se	$\hat{\beta}$	se	$\hat{\beta}$	se	$\hat{\beta}$	se
β_0	-6.23	0.04	-6.22	0.04	-6.23	0.05	-6.23	0.03
β_{age}	0.06	0.001	0.06	0.001	0.06	0.002	0.06	0.002
β_{sex}	-0.17	0.03	-0.17	0.03	-0.17	0.03	-0.17	0.04
β_M	0.28	0.13	0.26	0.13	0.29	0.14	0.28	0.11
β_{PI}	0.07	0.22	0.32	0.21	0.06	0.28	0.05	0.23
β_{nq}	0.96	0.04	0.95	0.04	0.97	0.05	0.96	0.02
β_{trade}	-0.16	0.05	-0.17	0.05	-0.15	0.07	-0.16	0.05
β_{tert}	-0.84	0.10	-0.83	0.10	-0.81	0.12	-0.81	0.11
β_{am}	-0.02	0.003	-0.02	0.003	-0.02	0.005	-0.02	0.005
β_{api}	-0.01	0.01	-0.01	0.01	-0.01	0.01	-0.01	0.01
β_{msex}	-0.22	0.09	-0.19	0.09	-0.22	0.09	-0.23	0.09
β_{pisex}	0.10	0.16	-0.03	0.16	0.13	0.15	0.13	0.16
β_{manq}	-0.20	0.13	-0.19	0.13	-0.23	0.14	-0.19	0.13
β_{mapsec}	0.05	0.17	0.08	0.17	0.01	0.18	0.00	0.15
β_{pinq}	-0.63	0.23	-0.81	0.22	-0.65	0.28	-0.53	0.20
β_{pipsec}	0.54	0.27	0.50	0.25	0.47	0.26	0.62	0.29

Table 5. Sensitivity of parameter estimates ($\hat{\beta}$) and standard errors (se), obtained from fitting the main effects logistic regression model to multiply-imputed synthetic data, to the imputation model.

	Imputation Model							
	Simple GLM		Complex GLM		Simple HB		Complex HB	
	$\hat{\beta}$	se	$\hat{\beta}$	se	$\hat{\beta}$	se	$\hat{\beta}$	se
β_0	-6.17	0.03	-6.18	0.04	-6.18	0.04	-6.18	0.03
β_{age}	0.06	0.001	0.06	0.002	0.06	0.002	0.06	0.001
β_{sex}	-0.18	0.02	-0.18	0.02	-0.19	0.02	-0.18	0.02
β_{M}	-0.06	0.06	-0.00	0.05	-0.00	0.05	0.01	0.04
β_{PI}	-0.35	0.06	-0.30	0.09	-0.30	0.07	-0.29	0.11
β_{niq}	-0.94	0.04	0.93	0.05	0.93	0.04	0.92	0.03
β_{trade}	-0.17	0.05	-0.14	0.05	-0.15	0.05	-0.17	0.04
β_{tert}	-0.91	0.08	-0.83	0.08	-0.83	0.07	-0.84	0.06

Table Five contrasts parameter estimates obtained from fitting the main effects logistic model to multiply imputed synthetic data generated under the simple and complex log-linear structures, for both the conventional (GLM) and hierarchical (HB) models. Differences between parameter estimates obtained under the simple and complex imputation models are smaller for the hierarchical imputation models. However, differences are relatively small overall with the possible exception of the Pacific Island estimate under the GLM imputation models for which the difference in point estimates was 0.05 with standard errors of 0.06 based on the data generated under the simple imputation model and 0.09 for the data generated under the complex GLM imputation model.

More substantial differences were apparent for parameter estimates obtained from fitting the ethnicity interaction logistic model to synthetic data generated under the simple and complex imputation models. Results are detailed in Table 6. The most obvious differences are for estimates of interaction terms which are all shrunk to be close to zero under the simple conventional log-linear imputation model (GLM). This reflects the absence of higher order interaction terms involving institutional care in the specification of the simple log-linear imputation structure and the reality that imputations from standard GLM models will exactly replicate the assumptions of the model. However results in Table 6 reveal that the impact of the log-linear model structure is considerably less when implemented within a hierarchical modelling set-up. While parameter estimates for the interaction terms are all pulled toward zero under the simple HB based imputation model, estimates for the Maori ethnicity by sex interaction (msex), Pacific ethnicity by no qualifications interaction (pinq), and to a lesser extent, the Maori ethnicity by no qualifications interaction (manq), all remain markedly non-zero. With the exception of the Pacific Island - post-secondary qualifications interaction term (pipsec), the logistic model interaction terms which are close to zero based on the simple HB based synthetic data correspond to interactions for which models fitted to the observed data suggest little support (see Table 4).

The fact that the estimates of the Pacific Island - post-secondary qualifications interaction term (pipsec) is shrunk very close to zero when estimated from the simple HB based synthetic data, suggests that, although there is some support for this effect in the data, (Table 4), a model omitting this effect nevertheless fits reasonably well so that, under an HB model omitting the Pacific-post secondary interaction (pipsec), the cell counts in the underlying table are shrunk towards a pattern of variation consistent with the absence of this interaction effect. Moreover, because of the small size of many of the Pacific cells, the model dominates in this region of the data and therefore posterior cell-specific estimates for the Pacific group will tend to shrink the observed data towards model based predictions more markedly than for other groups.

In addition to the sensitivity of the ethnicity interaction terms to the specification of the GLM imputation models, the estimated main effects for Maori (β_M) and Pacific

ethnicity (β_{PI}) also differed markedly between ethnicity interaction models fitted to the simple and complex GLM based synthetic data (Table 6). The corresponding differences for the simple and complex HB based synthetic datasets were considerably smaller. Thus it appears that synthetic data generated from HB based log-linear imputation models are more robust to the specification of the log-linear structural model than are conventional non-hierarchical versions of these models. This is entirely consistent with the theory of hierarchical modelling and the compromise nature of HB estimates, discussed in section 4.2

Table 6. Sensitivity of parameter estimates ($\hat{\beta}$) and standard errors (se), obtained from fitting the ethnicity interaction logistic model to multiply imputed synthetic data, to the imputation model.

	Imputation Model							
	Simple GLM		Complex GLM		Simple-HB		Complex-HB	
	$\hat{\beta}$	se	$\hat{\beta}$	se	$\hat{\beta}$	se	$\hat{\beta}$	se
β_0	-6.17	0.04	-6.23	0.05	-6.21	0.03	-6.23	0.03
β_{age}	0.06	0.001	0.06	0.002	0.06	0.002	0.06	0.002
β_{sex}	-0.18	0.02	-0.17	0.03	-0.17	0.02	-0.17	0.04
β_M	-0.07	0.07	0.29	0.14	0.17	0.11	0.28	0.11
β_{PI}	-0.40	0.27	0.06	0.28	-0.10	0.16	0.05	0.23
β_{nq}	-0.94	0.04	0.97	0.05	0.94	0.04	0.96	0.02
β_{trade}	-0.17	0.05	-0.15	0.07	-0.15	0.05	-0.16	0.05
β_{tert}	-0.91	0.08	-0.81	0.12	-0.81	0.07	-0.81	0.11
β_{am}	0.00	0.0003	-0.02	0.005	-0.01	0.004	-0.02	0.005
β_{api}	0.00	0.006	-0.01	0.01	-0.00	0.005	-0.01	0.01
β_{msex}	0.00	0.09	-0.22	0.09	-0.18	0.07	-0.23	0.09
β_{pisex}	0.00	0.11	0.13	0.15	-0.03	0.15	0.13	0.16
β_{manq}	0.01	0.14	-0.23	0.14	-0.09	0.11	-0.19	0.13
β_{mapsec}	-0.01	0.19	0.01	0.18	0.01	0.15	0.00	0.15
β_{pinq}	0.05	0.27	-0.65	0.28	-0.24	0.12	-0.53	0.20
β_{pipsec}	-0.02	0.40	0.47	0.26	0.02	0.07	0.62	0.29

4.4.2 Impact of synthetic data on confidentiality

Methods for evaluating disclosure risks associated with multiply imputed synthetic data are not well developed. Nevertheless it seems reasonable to assume that the synthetic nature of the data, coupled with the fact that multiple datasets are produced should discourage most attempts at using multiply imputed synthetic datasets to identify individuals. However, with data representable via a multiway contingency table, as is the

case for the institutional care data, a naive analyst may try to infer true cell counts from averages over the multiple imputed synthetic datasets. Such an analyst would be naive because the smoothing implicit in the models underpinning the imputation process ensures that average imputed cell counts will often deviate somewhat from the actual cell counts. The degree of variability in the imputed cell counts is also likely to influence an analyst's confidence in their ability to infer the observed cell-counts from the multiple synthetic datasets. In an extreme case, a cell for which the synthetic datasets yielded a count of one on each imputation, could very well be assumed by an analyst to correspond to a cell for which the true count was one. If, in fact, that cell did have a true cell count of one then this would constitute a disclosure of an individual record. If not, then the analyst may still act on their erroneous belief that they have identified a unique individual. Whether this would harm anyone except the analyst is unclear and would depend on the intended use for the erroneous information.

In order to explore these issues in the context of the institutional care data we compared average imputed cell counts with actual counts and, for sensitive cells, particularly those for which the observed count was one, examined the distribution of counts over the multiple imputed synthetic datasets.

Figures 1 and 2 compare average synthetic cell counts generated under conventional (panel B) and HB (panel C) log linear imputation models with the observed counts. Average imputed cell counts were rounded to the nearest integer. Figure 1 shows average imputed counts under the complex log-linear imputation models, while Figure 2 gives the analogous results for the conventional and hierarchical versions of the simple log-linear imputation models. As a reference point, randomly rounded cell counts are also compared with the observed counts (panel A). Because of the variation in the magnitude of the cell counts a \log_{10} scale is used to display the counts. Cell counts of zero were shifted to 0.1 and therefore show as -1 in the figures. Similarly cell counts of 1 show as 0 in the figures. The average imputed cell counts appear to exhibit more variation from the observed counts than is the case for randomly rounded counts. Therefore, if the degree of variation from true counts exhibited by the randomly rounded counts is regarded as sufficient to protect confidentiality then the same must be said for the average synthetic counts. One potential advantage of the random rounding is that true counts of exactly 1 are shifted to either 0 or 3. However, inspection of the average imputed counts reveal that under all imputation models considered there is sufficient variability both in the average imputed counts corresponding to cells with a true count of one and in the true counts corresponding to average imputed counts of of one, to warrant an official statistics agency advising that "it cannot be assumed that average imputed cell counts correspond to the actual counts, and in particular, average imputed cell counts of 1 cannot be assumed to indicate true cell count of 1".

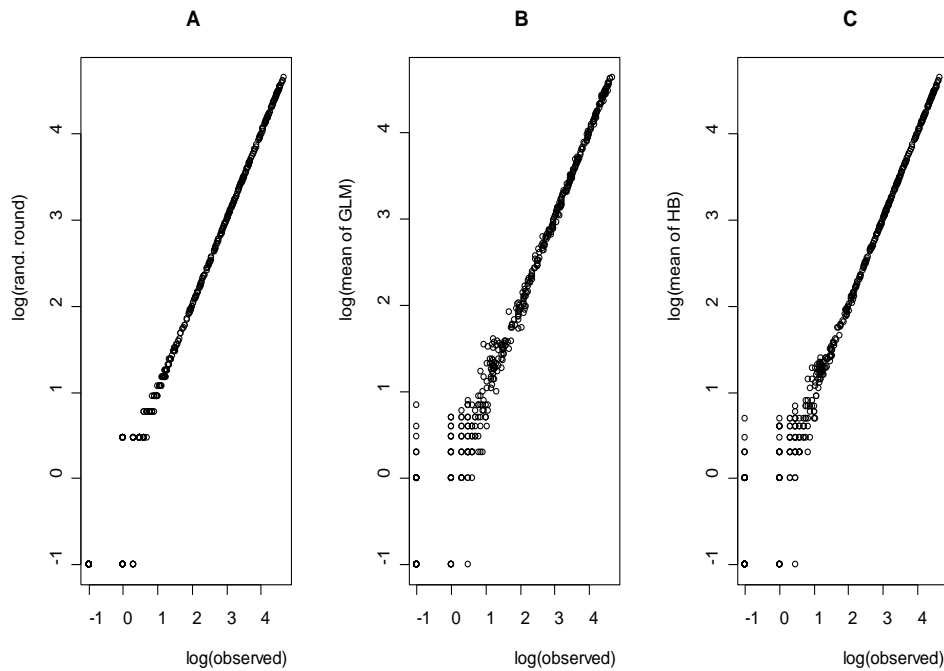


Figure 1: Comparison of randomly rounded cell counts (A) and average synthetic cell counts under complex GLM (B) and HB (C) models with observed cell counts

In terms of the distribution of cell counts across multiple synthetic datasets, situations in which a high proportion of counts for a particular cell were equal to one, would be of some concern, particularly if the cell in question had a true count of one. For each of the 28 cells in the institutional care data which had a true count of one we examined the distribution of cell counts over the 100 imputed synthetic datasets, generated under each of the four imputation models (conventional GLM and HB versions of both complex and simple log-linear structures). The extreme situation of synthetic cell counts uniformly equal to one did not arise under any of the imputation models. There was a group of six cells for which the maximum synthetic cell count was 1 or 2, under each of the imputation models. However in these cases the median cell count across the multiple imputations was 0, the mean was close to zero and in the majority of cases the upper quartile of the distribution was 0. Therefore for these six low-variability cells an analyst endeavouring to infer the true cell count from the distribution of multiply imputed counts would be more likely to infer, incorrectly, that the true cell count was zero, than that the true count was, in fact, equal to 1. On the other hand, among the 28 cells with a true count of one, the number of the cells with a median imputed count of one ranged from 6, under the complex HB imputation model to 3, under the simple non-hierarchical GLM model. Under all imputation models, cells with a median imputed count of one all had minimum values of zero and maxima of at least four, with upper quartiles equal to 1 or 2. This level of variability across imputations should

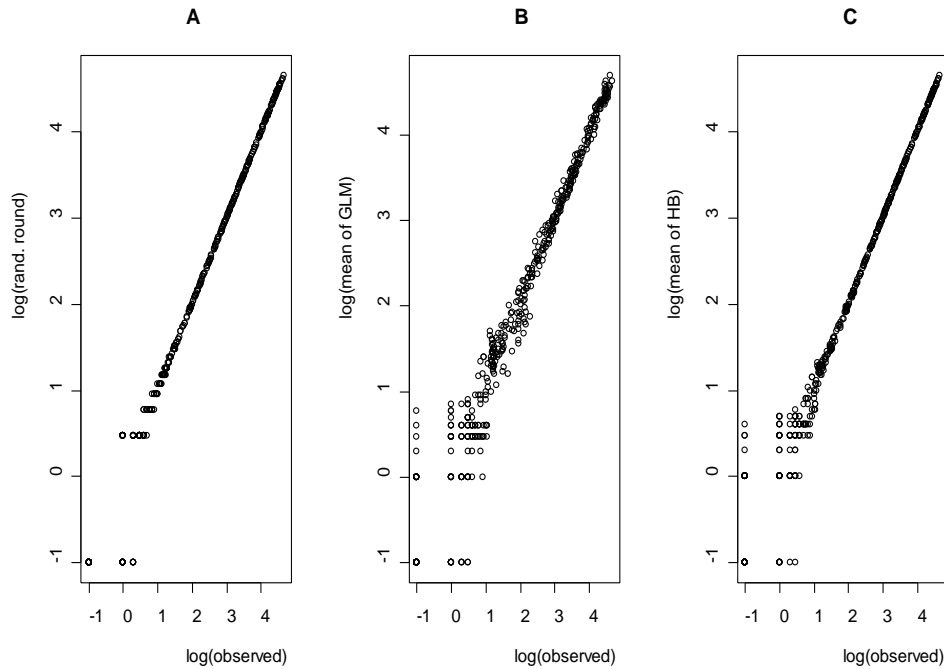


Figure 2: Comparison of randomly rounded cell counts (A) and synthetic cell counts under simple GLM (B) and HB (C) models, with observed cell counts

discourage analysts from being confident as to the true cell count.

5 Discussion

In this paper we have investigated the multiple imputation approach to creating synthetic datasets, originally proposed by Rubin (1993). In keeping with the theory underlying multiple imputation we emphasised the role of Bayesian predictive inference in providing a conceptual framework, both for understanding the problem of creating synthetic datasets and for generating possible solutions to the problem. The Bayesian paradigm also underpins methods for the analysis of multiply imputed synthetic data, which involves combining inferences across the generated synthetic datasets, as summarised in section 3.4. The combining rules derived by Raghunathan et al (2003) can be viewed as providing a large sample approximation to the posterior distribution of the quantity of interest, conditional on the synthetic data.

We illustrated the multiple imputation approach to the generation and analysis of synthetic data via application to a dataset containing information on social variation in the use of institutional care. Although this example is of unrealistically small dimension and simplified because all five variables can be regarded as categorical it nevertheless serves to illustrate several important features of the multiple imputation approach to synthetic data. Firstly, inferences obtained from multiply imputed synthetic data can

be similar to those obtained from observed data, particularly when the analytical model is of a simple form (see Table 2). However, inferences obtained from synthetic data are potentially sensitive to the choice of imputation model (Table 6). In this regard, fitting imputation models within a hierarchical Bayesian modelling framework has considerable appeal because these models allow for model uncertainty and produce estimates which are compromises between raw data and model-based estimates and therefore lead to reduced sensitivity of inferences to model specification.

Synthetic data released by a data collection agency will reflect both the observed data and the prior knowledge of the agency. Here, we are interpreting prior knowledge in a broad sense to include situations in which the data collector has access to variables which would never be released to external analysts but which can be used to construct imputations, as well as knowledge of non-response mechanisms, results of previous studies and general contextual knowledge possibly including ideas of smoothness and similarity between groups. Thus the analyst of synthetic data benefits from the imputer's prior knowledge, provided the imputer is able to fairly represent this knowledge in constructing the imputations. A related point is that the evaluation of the validity of analyses based on synthetic data should not focus solely on whether inferences based on synthetic data are similar to those based on the observed data. When the imputer's prior information is accurate, analyses of the synthetic data may yield more accurate inferences than would analyses of the observed data undertaken in the absence of the imputer's prior information. This issue is particularly relevant to inference for small sub-groups of the population for which estimates from survey data alone may be unstable, an issue now receiving considerable methodological attention under the heading of "small-area estimation" (Rao, 2003). Evaluation of the validity of synthetic data based inferences should, ultimately, be concerned with the accuracy of inferences for *population* parameters, rather than focussing solely on replication, of *sample* estimates. Simulation studies such as those reported in Reiter (2005) will be required to address this issue. However, such evaluations lay outside the scope of current project.

In terms of confidentiality protection, our results are somewhat tentative being based on a single example and given the current lack of standard measures for evaluating disclosure risk for multiply imputed synthetic datasets. However, we note that even when the multiply imputed institutional care datasets were reduced to a single summary dataset by averaging, the deviations in average cell counts from the corresponding observed values appears slightly greater than was achieved by random rounding, which is an accepted technique for confidentialising data which can be represented in tabular form. Thus, the confidentiality protection afforded by multiply imputed synthetic data may be greater than is provided by current techniques. This conjecture seems even plausible when due regard is given to fact that under the MI approach, multiple

versions of the synthetic data are created. This should discourage most rational people from using the data to identify individuals. However it is clear that specific measures of disclosure risk, tailored to the particular features of synthetic data remain to be developed.

Although average imputed cell counts deviated more from actual counts than did randomly rounded cell counts, parameter estimates obtained from fitting logistic regression models to multiply imputed synthetic data, were closer to the those obtained from the original data than were estimates obtained from fitting randomly rounded data. This reflects the fact the fact that random rounding is an example of a confidentialising technique which adds random noise and therefore perturbs observed data in a random fashion. In contrast, model based approaches perturb observed data by moving the observed data towards some pattern of variation which is typically less complex than exhibited by the raw data. In essence, modelling induces confidentiality protection by smoothing towards some assumed plausible pattern of variation and therefore can be regarded as reducing noise rather than adding noise. In the case of HB modelling, the dataset itself largely determines the amount of smoothing as a function of overall model-fit and the amount of information contributed by each cell. A poor fitting prior model will not induce much shrinkage of observed cell counts towards the counts predicted under the model. On the other hand, the degree of shrinkage is also a function of cell size, with shrinkage towards the prior model tending to be greater for smaller cells than for larger cells.

Provided the analyst's model is simpler than the imputer's model the smoothing induced by the imputation model is unlikely to have much effect on the analyst's inferences because the additional complexity of the imputer's model implies a lesser degree of smoothing than is assumed by the analyst's model. This is illustrated in Table 3 which shows very similar parameter estimates and standard errors for a simple "main effects" logistic model fitted to the original institutional care data and to the multiply imputed data. However, as the complexity of the analyst's model increases so does the potential for sensitivity to the specification of the imputation model. This is because more complex aspects of variation which may be omitted from the imputer's model may in fact be relevant to more complex analytical models. Of course, if there is little support in the data for these higher-order associations then the imputer's model will not harm inference. The preceding discussion points to the need for imputation modelling strategies which provide some degree of robustness to model specification such as hierarchical Bayesian modelling which explicitly allows for model uncertainty. Bayesian model averaging may also provide a workable approach to protecting multiply imputed synthetic data from model mis-specification (Madigan & Raftery, 1994).

In realistically high-dimensional problems imputation modelling will proceed by fitting a sequence of conditional models (Reiter 2005). In principle each of these models

could be specified and fitted within the hierarchical modelling paradigm. For example given an initial set of categorical variables, modelled via a hierarchical Poisson log-linear model a second set of continuous variables could, after suitable transformation, be modelled conditionally on the categorical variables via a hierarchical multivariate normal model (Everson & Morris 2000). As with the hierarchical Poisson model the computational demands of model fitting via Markov Chain Monte Carlo can be avoided for such models (Everson & Morris, 2000). However the practical limits of the hierarchical Bayes approach in terms of the dimension of the problems which can be handled remains to be determined.

Among other possible imputation modelling paradigms, Bayesian network modelling holds promise (Cowell et al.1999, Di Zio et al 2004). Bayesian network models of multivariate data exploit conditional independencies to model joint distributions via parsimonious sequences of conditional models. Software implementations of Bayesian networks typically incorporate automated model search procedures, which may be useful in identifying imputation models (Bottcher & Dethlefsen 2003). However, typical implementations appear ultimately to condition on a selected model which may induce an unwarranted degree of model dependence. Bayesian model averaging (Madigan & Raftery, 1994) could however be applied to Bayesian networks, effectively yielding posterior inferences based on a mixture of Bayesian network models, thereby decreasing model dependence. Recent developments in nonparametric Bayesian modelling (Walker et al 1999) may also provide a useful source of approaches for developing imputation models. While these approaches greatly reduce the requirement to specify specific parametric forms for data models, it is unclear whether they have sufficient flexibility to model complex data structures, such as those with obvious hierarchical structure as is often apparent in household surveys. In contrast hierarchical models handle such structures in a very natural manner. The simplest non-parametric Bayesian approach for generating synthetic datasets is the Bayesian bootstrap (Rubin, 1981). However, because this procedure effectively resamples complete records from the observed dataset it is not desirable from a confidentiality viewpoint. A user of a synthetic data generated by a Bayesian bootstrap procedure can be sure that each record they see in the synthetic datasets, corresponds to a real record in the observed data. More recently developed non-parametric Bayesian procedures (Walker et al 1999) potentially offer more confidentiality protection than the Bayesian bootstrap.

One potential objection to the multiple imputation paradigm for constructing synthetic data is the perception that handling and analysing multiple datasets will be too difficult for users. However in an age of jackknifing, bootstrapping, replicate weights, Monte Carlo and Markov Chain Monte Carlo, arguing that repeated analysis is beyond the reaches of most statistical analysts is not a credible position. Rubin's rebuttal of similar objections to the multiple imputation paradigm for conventional missing data

problems effectively dismantles the objection as it applies to most datasets (Rubin, 1996). With extremely large datasets there may be practical issues in terms of the processing time required to run multiple analyses and investigation of the number of imputations required for such large datasets may warrant further investigation

The multiple imputation paradigm provides a promising approach for the development of practical methods for creating synthetic datasets. An important advantage over ad-hoc approaches to producing confidentialised datasets is that the multiple imputation approach is embedded within a theory of inference. It is therefore clear to potential users how they should proceed to draw valid inferences from multiply-imputed synthetic data. The same cannot be said of data confidentialised by a variety of ad-hoc methods. From the data-collector's perspective the MI approach to synthetic data has two distinct advantages. Firstly, with appropriate communication of the synthetic nature of the data and the fact that multiple versions of the data are released, most users should be convinced that identification of specific individuals from the released data would be an extremely difficult task. Secondly the MI approach allows for a synthetic dataset to be constructed in a modular fashion. If after an initial release further variables are required to be released, the only additional modelling required is to model the new variables conditionally on the previously released variables. This model can then be used to impute the new variables conditionally on the actual values of the previously released variables. In this way consistency can be maintained between sequential releases of a synthetic dataset.

While many practical details of the multiple imputation approach to synthetic data remain to be worked out the potential gains in terms of data access and valid inference are sufficiently substantial to suggest that such investigations will yield a substantial benefit, both to data-suppliers and to the research and policy communities.

References

- ABOWD, J.M. & WOODCOCK, S.D. (2001) Disclosure limitation in longitudinal linked data. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. P. Doyle, J. Lane, J. Theeuwes & L. Zayatz eds. Amsterdam: North-Holland.
- AGRESTI, A. (1990) *Categorical data analysis* New York: Wiley
- ALBERT, J.H. (1988) Computational methods using a Bayesian hierarchical generalised linear model. *J. Amer. Statist. Assoc.* **83**, 1037-1044.
- BETHLEHEM, J.G., KELLER, W.J. & PANNEKOEK, J. (1990) Disclosure control of microdata. *J. Amer. Statist. Assoc.* **85**, 38-45
- BOTTCHEER, S.G. & DETHLEFSEN C. (2003) deal: A package for learning Bayesian networks. *Journal of Statistical Software* **8**(20)
- CHRISTIANSEN, C.L. & MORRIS C. (1997). Hierarchical Poisson regression modelling. *J. Amer. Statist. Assoc.* **92**, 618-632
- COWELL, R.G., DAWID, A.P., LAURITZEN, S.L. & SPIEGELHALTER, D.J. (1999) *Probabilistic Networks and Expert Systems*. New York: Springer-Verlag.
- DANIELS, M.J. (1999). A prior for the variance in hierarchical models. *Canadian Journal of Statistics* **27**, 567-578.
- DI ZIO, M. SCANU, M., COPPOLA, L. LUZI, O. & PONTI, A. (2004) Bayesian networks for imputation. *J. Roy. Statist. Soc. A.* **167**, 309-322
- DOYLE, P., LANE, J.I., THEEUWES, J., ZAYATZ, L.V. (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam: North Holland
- DUNCAN, G.T., JABINE, T.B. & DE WOLF, V.A. (1993) *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Washington, D.C.: Academy Press.
- EVERSON, P.J. & MORRIS, C.N. (2000). Inference for multivariate normal hierarchical models. *J. Royal Statistical Society B* **62**, 399-412.
- FIENBERG, S.E. & MAKOV, U. (1998) Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics* **14**, 385-397.
- FIENBERG, S.E. & MAKOV, U. (2001). Uniqueness, urn models and disclosure risk. *Research in Official Statistics* **4**, 23-40.

- FRANCONI, L. & STANDER, J. (2003) Spatial and non-spatial model-based protection procedures for the release of business microdata. *Statistical Computing* **13**, 295-306.
- GREENLAND, S. (1995) Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology* **6**, 356-365
- HILL S., ATKINSON, J. & BLAKELY T (2002). *Anonymous record linkage of census and mortality records: 1981, 1986, 1991, 1996 census cohorts*. Wellington, NZ: Department of Public Health, Wellington School of Medicine and Health Sciences, University of Otago
- KENNICKEL, A.B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances, in *Record Linkage Techniques 1997*. W. Alvey & B. Jamerson eds., pp 248-267. Washington D.C.: National Academy Press.
- LIU, F. & LITTLE, R.J.A. (2002) Selective multiple imputation of keys for statistical disclosure control in microdata. In *Proc. Joint Statistical Meetings*, pp 2133-2138, Blacksburg: American Statistical Association.
- LO, A. Y. (1988) A Bayesian bootstrap for a finite population *Annals of Statistics* **16**, 1684-1685.
- MADIGAN, D.M. & RAFTERY, A.(1994) Model uncertainty and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89**, 1535-1546.
- MEEDEN, G., VARDEMAN, S. (1991) A noninformative Bayesian approach to interval estimation in finite population sampling. *J. Amer. Statist. Assoc.* **86**, 972-980.
- MENG, X.L. (1994) Multiple imputation with uncongenial sources of input (with discussion). *Statistical Science* **9**, 538-574.
- POLETTINI, S. & STANDER, J. (2004) A Bayesian hierarchical model approach to risk estimation in statistical disclosure limitation. In *Privacy in Statistical Databases 2004*, Domingo-Ferrer, J. & Torra. V., eds., pp247-261, Berlin: Springer-Verlag.
- R Development Core Team (2003) *R: A language and environment for statistical computing*. Vienna, Austria: R foundation for statistical computing. URL <http://www.R.project.org>
- RAGHUNATHAN, T.E. REITER, J.P. & RUBIN, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* **19**, 1-16.
- RAO, J.N.K. On variance estimation with imputed survey data. (1996) *J. Amer. Statist. Assoc.* **91**, 499-506
- RAO, J.N.K. (2003) *Small Area Estimation*. New York: Wiley

- REITER, J. P. (2002) Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* **18**, 531-543
- REITER, J. P. (2005) Releasing multiply imputed. synthetic public use microdata: An illustration and empirical study.. *J. Roy. Statist. Soc. A.* **168**, 185-205.
- RUBIN, D.B. (1981) The Bayesian bootstrap. *Annals of Statistics* **9**, 130-134
- RUBIN, D.B. (1987) *Multiple imputation for non-response in surveys*. New York: Wiley
- RUBIN, D.B. (1993) Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462-468
- RUBIN, D.B. (1996) Multiple imputation after 18+ years. *J. Amer. Statist. Assoc.* **91**, 434-489
- SAMUELS, S.M. (1998) A Bayesian species-sampling-inspired approach to the unique problem in microdata disclosure risk assessment. *Journal of Official Statistics* **10**, 31-51
- SKINNER, C.J. MARSH, C., OPENSHAW, S. & WYMER, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics* **10**, 31-51
- SKINNER, C.J. & HOLMES, D.J. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, **14**, 3761-372
- WALKER, S.G., DAMIEN, P., LAUD, P.W. & SMITH A.F.M.(1999) Bayesian non-parametric inference for random distributions and related functions. *J. Roy. Statist. Soc. B* **61**, 485-527
- WILLENBORG, L. & DEWAAL, T. (1996). *Statistical Disclosure Control in Practice*. New York: Springer.