Journal of Official Statistics, Vol. 25, No. 4, 2009, pp. 549–567

Using Bayesian Networks to Create Synthetic Data

Jim Young[1], *Patrick Graham*[2], *and Richard Penny*[3]

A Bayesian network is a graphical model of the joint probability distribution for a set of variables. A Bayesian network could be used to create multiple synthetic data sets that are then released by an official statistics agency while the original data remain confidential, so that an analyst outside the agency can explore associations between an attribute of interest and other variables. The process is illustrated with an example. Inferences from the original data are compared to inferences from synthetic data created by a single Bayesian network and by Bayesian model averaging over a set of networks. Informative prior information is needed in order to assign appropriate weights to each network in this set if synthetic data are to have both good inferential properties and an acceptable risk of disclosure. This sensitivity to prior information will make it difficult for an official statistics agency to use Bayesian networks to automate the process of creating synthetic data.

Key words: Confidentiality; hierarchical Bayesian modelling; multiple imputation.

1. Introduction

A Bayesian network is a graphical model of the joint probability distribution for a set of variables. The model has two components: a graphical structure and a set of conditional probability distributions. Such a model is sometimes called a belief network. A Bayesian network is typically used for probabilistic inference about one variable in the network given the values of other variables. Ramoni and Sebastiani (2001) construct a Bayesian network to summarise relationships between the variables collected in a household survey. They suggest that users outside an official statistics agency could access this Bayesian network via the internet and carry out their own inference without having direct access to the original data, thus preserving data confidentiality.

Bayesian networks have also been used to impute missing survey data (Di Zio et al. 2004). These authors then suggest that a Bayesian network could be used to create a synthetic data set; Thibaudeau and Winkler (2002) make a similar suggestion. However, inference from synthetic data requires multiple synthetic data sets. These synthetic data

[1] Vital Statistics Ltd, 85b Barrington Street, 8024 Christchurch, New Zealand. Email: jyoung@actrix.co.nz
[2] Christchurch School of Medicine and Health Sciences, University of Otago, PO Box 4345, 8011 Christchurch, New Zealand. Email: patrick.graham@otago.ac.nz
[3] Statistics New Zealand, Private Bag 4741, 8140 Christchurch, New Zealand. Email: richard.penny@stats.govt.nz
Acknowledgments: This research was funded by Statistics New Zealand through the Official Statistics Research (OS Research) fund. This article uses material from a report prepared for OS Research (Young 2008a); Statistics New Zealand holds copyright to the report and to the research. The views expressed in the article are those of the authors and do not represent an official view of Statistics New Zealand. We thank Kim Cullen (OS Research, Statistics New Zealand) for facilitating this research and Claus Dethlefsen (Centre of Cardiovascular Research, Aarhus University Hospital, Denmark) for his help with "deal," a Bayesian network package for R.

sets could be released for wider use while the original data remain confidential. The method is an extension of multiple imputation, and was developed by Rubin (1993); Reiter (2002) and Raghunathan et al. (2003). Reiter (2005) uses a sequence of generalised linear models to create an appropriate multiple imputation model. Graham and Penny (2007) take a hierarchical Bayesian approach that allows for uncertainty in a generalised linear model specified *a priori* for the original data (Albert 1988; Christiansen and Morris 1997) so that their synthetic data are robust to misspecification of this generalised linear model. They note that while a Bayesian network could be used as a multiple imputation model, synthetic data from a network would be conditional on the network selected. They suggest using Bayesian model averaging to account for uncertainty in the selection process so that the resulting synthetic data are robust to network mis-specification (Graham and Penny 2007, p. 41). Madigan and Raftery (1994) consider Bayesian model averaging in the context of graphical models.

A Bayesian network can be seen as an automated method of "learning" about probabilistic relationships in data (Heckerman et al. 1995; Ramoni and Sebastiani 2001). An automated method of creating synthetic data – that is, one that requires little information from the user – would clearly appeal to an official statistics agency (see comments in both Ramoni and Sebastiani 2001; Thibaudeau and Winkler 2002). But although a number of authors have suggested using a Bayesian network to create synthetic data (Thibaudeau and Winkler 2002; Di Zio et al. 2004; Graham and Penny 2007), the approach does not seem to have been tried.

2. An Example

We illustrate a Bayesian network approach to creating synthetic data using the institutional care data set of Graham and Penny (2007) and Graham et al. (2009). The data consist of five variables: age (in 10 five-year categories), sex, ethnicity (two minorities, Maori and Pacific Islanders, and the non-Maori non-Pacific Islanders majority), highest educational qualification (none, secondary and post-secondary) and an indicator for permanent residence in a health care institution. Note that for highest educational qualification, we combine trade and tertiary qualifications into a single post-secondary school category, whereas Graham et al. (2009) use all four categories. Hence by cross-classifying all variables, these data can be represented by a table of frequency counts with 360 cells.

The institutional care data set is challenging from both analytic and confidentiality perspectives. The prevalence of institutional care in these data is only 0.3%. This and the uneven distribution of both ethnicity and highest educational qualification lead to many cells with small counts. Ethnic groups differ in size from 1,664,481 for the non-Maori non-Pacific Islanders majority to 219,126 for Maori and 74,841 for Pacific Islanders. The distribution of highest educational qualification varies with ethnicity: 41% of the non-Maori non-Pacific Islanders majority have a post-secondary school qualification but only 25% and 21% of Maori and Pacific Islanders respectively are in this category. The number in institutional care is low in many subgroups: when summed over the 20 categories of age and sex, only 76 Maori and 31 Pacific Islanders with post-secondary school qualification are in institutional care. Within the 360 cell cross-classification, 22 cells have a frequency

of one and 19 cells have a frequency of zero. Graham and Penny (2007, pp. 18–20) describe how these data were collected.

We envisage a scenario in which a researcher (Reiter's "analyst") requests a data set with a clear analytic focus (here institutional care) and an official statistics agency (Reiter's "imputer") wishes to release synthetic data suitable for the analyst's purpose (Reiter 2005, p. 187). We assume the imputer would want to create synthetic data that give almost the same estimates as the real data for any logistic model for institutional care whose complexity is less than a model with all associations between institutional care and any pair of predictor variables. If the imputer achieves this, the synthetic data will be released together with a statement advising the analyst of this limitation. Reiter notes that the imputer needs to release information to help the analyst decide whether the synthetic data are suitable for their analysis (Reiter 2005, p. 189).

We use two logistic regression models to assess the impact of synthetic data on inference. The first model – the "main effects" model – has institutional care as the response and all other variables as predictors without any interaction terms. The second model – the "one way interaction" model – has additional terms representing all one way interactions between every pair of predictor variables. In both models, age is treated as a continuous variable. Having fit a Bayesian network model to the institutional care data set, we use the fitted network as a multiple imputation model and generate 100 synthetic data sets each of the same size as the original data set (approximately 2 million individuals). Both main effect and one way interaction logistic models are fit to each of the synthetic data sets in turn. Overall estimates for logistic model parameters and their standard errors are then calculated from the 100 estimates, one for each synthetic data set, using the combining rules given by Reiter (2002). Finally these overall estimates from synthetic data are compared to estimates the analyst would have obtained had the analyst been able to fit these two logistic regression models to the real data.

Synthetic data reduce but do not eliminate the risk of disclosure, and whether this risk is acceptable or not to an official statistics agency depends on context. To assess the relative risk of disclosure, we provide plots of the median synthetic cell count versus the real cell count. An analyst might be tempted to use the summary statistics of synthetic data to predict the number of individuals in the real data that belong to a particular subgroup, and this sort of predictive disclosure is less likely if real counts are more variable for any given median synthetic count. In addition, we give the percentage of cells with a real count of one among cells with a median synthetic count of one. This is a synthetic data equivalent to the risk of inferring uniqueness in a population given uniqueness in a sample (see, for example Fienberg and Makov 1998). Hence plots showing the variability of real cell counts for a given median synthetic cell count and the percentage of real cell counts of one in cells with a median synthetic count of one are both relative measures with which we can compare the disclosure risk inherent in synthetic data from different imputation models.

3. Bayesian Networks

A Bayesian network consists of a graphical structure and a set of conditional probability distributions. The graphical structure is a set of nodes, each node representing a discrete or continuous variable, and a set of arrows between nodes. Probabilistic inference from

a Bayesian network requires that arrows from one node do not lead back to that same node (Jensen 2001, p. 19) and so if feedback is a feature of the variables being modelled, then a sequence of nodes can be used to represent the same variable at different times (see Neapolitan 2004, pp. 265–269). In the language of Bayesian networks, if there is an arrow from one node to another, the former is the "parent" node and the latter the "child." Associated with each node is a conditional probability distribution. This gives the probability of each value of a node given the values of its parents. The absence of an arrow between two nodes implies conditional independence; that is, the two variables represented by these nodes are independent given knowledge of the values of their parents.

When constructing a Bayesian network, it is usually easier to specify "root causes" first and then the variables they influence and so on. Connecting each node in this way ensures the network is acyclic and should lead to a simpler network, with fewer arrows and with conditional probability distributions that are easier to think about and assign values to (Russell and Norvig 1995, pp. 441–443). However, the arrows in a Bayesian network do not imply causality. A Bayesian network is a model for an overall joint probability distribution for the set of variables, and several networks with arrows in different directions may imply the same set of conditional independence relationships (Cowell et al. 1999, pp. 252–253).

A Bayesian network is typically used for probabilistic inference about one variable in the network given the values of other variables. The usual rules of probability are applied to the set of conditional probability distributions, one for each node. In particular, inference makes use of Bayes' rule (see Cowell et al. 1999, p. 15) and this gives Bayesian networks its name, rather than any commitment to Bayesian methods. In practice, constructing a Bayesian network requires selecting an appropriate structure for the network and then estimating its parameters. Current methods attempt to carry out both tasks concurrently, rather than simply estimating the parameters of a network asserted by a subject-matter expert. Network selection has two sub-tasks (Cowell et al. 1999, pp. 243–263): searching for suitable networks and evaluating the various networks found. A variety of methods of searching and evaluation have been proposed (Tsamardinos et al. 2006).

4. Synthetic Data from a Single Network

4.1. Methods

Bottcher and Dethlefsen (2003) have written Bayesian network software that uses the R system for statistical computing and graphics (R Development Core Team 2004). Their software, "deal," is a Bayesian implementation of a Bayesian network. That is, the user must specify a prior network and prior conditional distributions for its nodes; observed data are then used to update both the network and its conditional distributions. For a discrete node, the observed data are represented by a set of multinomial conditional probabilities and the prior distribution for these conditional probabilities is Dirichlet, as is usual in a Bayesian implementation of a Bayesian network (see Heckerman et al. 1995; Cowell et al. 1999, p. 193).

Although this software allows continuous nodes, all the networks in our example consist entirely of discrete nodes. There are two reasons for this. First, the example we use has a binary attribute as its focus. It is natural to think of this variable as at the end of a chain of influence. Exact probabilistic inference is only possible if normal distributions are used for continuous nodes and if a continuous parent does not have a discrete child (Jensen 2001, p. 69; Cowell et al. 1999, pp. 125–136). To avoid assigning continuous parents to our binary attribute, each continuous parent must be categorised into discrete intervals. Essentially this reduces the network from a mixture of continuous and discrete nodes to one that consists entirely of discrete nodes. Second, the code required to create synthetic data is much easier to adapt from code already available in "deal" if a network has only discrete nodes.

In "deal," the user specifies a prior network structure and prior conditional probability distributions for each of the structure's nodes. The user also specifies the size of an imaginary database that would give rise to this prior information, as a way of expressing confidence in this prior. Data are used to update the probability distribution for each node. Under certain assumptions, using the data and prior one can then calculate a posterior joint probability distribution for any other network structure for these nodes (Heckerman et al. 1995). A "greedy" search algorithm is used to generate alternative network structures and to select the structure with the highest posterior probability given data and prior.

In the absence of user-supplied prior information, certain defaults operate. These are: the prior network structure is one where all nodes are independent; prior conditional distributions for each node have equal probabilities for each category; and the size of the imaginary database is set to twice the inverse of the largest joint prior probability. (With all nodes independent and equal probabilities for each category of each node, this default for the size of the imaginary database is equivalent to two observations for each cell in a cross-classification of all discrete variables. Other noninformative defaults have been suggested – see Heckerman et al. (1995, p. 212).)

There is, however, no easy way to specify prior probabilities for different network structures. Hence the network selected in a "greedy" search has the highest posterior probability only if the user considers *a priori* that all network structures are equally likely. The user can provide some prior information about admissible network structures by specifying a "ban list." This is a list of arrows that the user is willing to rule out. In essence, this is attaching a prior probability of zero to any network with an inadmissible arrow. For example, where a binary attribute is at the end of a chain of influence, it makes sense to ban any arrow from this variable back to other nodes. The "greedy" search algorithm disregards any network structure containing one of these banned arrows and this reduces the search space. Using a "ban list," the user can easily specify an order among variables and have some control over the sequence of conditional distributions in the model selected to represent the joint distribution of all variables.

To fit a Bayesian network in "deal" to the institutional care data, we use the method described by Bottcher and Dethlefsen (2003). In brief: (1) we specify a prior consisting of a prior network, prior conditional probability distributions for each of its nodes, and our confidence in this prior information; (2) a joint posterior probability distribution is calculated for this network given the institutional care data and this prior; and (3) a search algorithm generates alternative networks and selects the one with the highest posterior

probability given the data and this prior. This network is then used to generate synthetic data (Young 2008a, Section 8.1). We consider a sequence of prior network models, starting with a model that requires no information from the imputer.

4.2. Results

The first network model fitted (Model A) starts with a completely unspecified network as its prior network, with no arrows between any of the five nodes (Figure 1 left). All defaults operate: within each node, equal prior probabilities are assumed for each category, and the imaginary database size is set at 720 (i.e., two observations for each cell in the cross-classified data). The "greedy" search algorithm leads to a network with institutional care (InCare) influenced by age, ethnicity (Ethn) and highest educational qualification (Educ) but not by sex (Figure 1 right). Inference using synthetic data from this network is discouraging. Even in a main effects model, estimates for synthetic data from this network differ appreciably from estimates for the real data for all parameters except age (Table 1).

In an effort to improve inference, we provide some prior information about network structure in Model B. We believe that sex and ethnicity might influence the age structure of this population but would not influence each other; all three variables might influence the highest educational qualification held; and all four variables might influence whether someone is in institutional care. We are then willing to rule out influence in the reverse direction and these considerations lead to the "ban list" shown in Figure 2 (left). With this prior network, the "greedy" search algorithm leads to a network with institutional care influenced by the same variables as before (Figure 2 right). The only real difference between Models A and B is the absence of an arrow between sex and ethnicity in B – all other differences are just in the direction of the arrows. Hence both Models A and B lead to similar inference when fitting a main effects logistic regression model to their respective synthetic data (Table 1).

The network score given in each figure is the sum of the log marginal probability of each observation given the network structure and prior information (see Kass and Raftery 1995, pp. 776–777; Cowell et al. 1999, pp. 248–249). The higher score in Figure 1 implies that

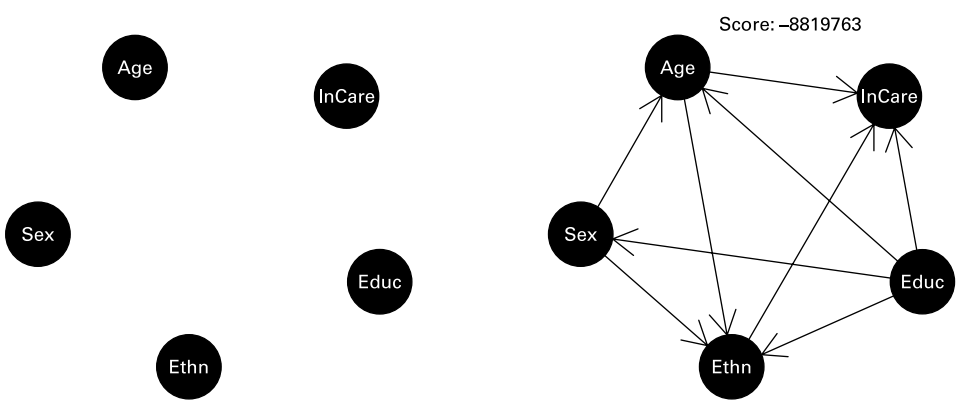

Fig. 1. Model A – its prior network (left) and the result of a "greedy" search (right) [Nodes: InCare, institutional care; Educ, highest educational qualification; Ethn, ethnicity]

Table 1. Parameter estimates and their standard errors for a main effects logistic regression model fit to the institutional care data and to synthetic data from Bayesian network Models A to E

	Real data		Model A		Model B		Model C		Model D		Model E	
	$\hat{\beta}$	$SE_{\hat{\beta}}$	$\hat{\beta}$	$SE_{\hat{\beta}}$	$\hat{\beta}$	$SE_{\hat{\beta}}$	$\hat{\beta}$	$SE_{\hat{\beta}}$	$\hat{\beta}$	$SE_{\hat{\beta}}$	$\hat{\beta}$	$SE_{\hat{\beta}}$
β_0	−8.15	0.07	−8.02	0.10	−8.01	0.09	−7.94	0.09	−8.16	0.11	−7.82	0.08
β_{age}[a]	0.61	0.01	0.60	0.02	0.60	0.01	0.60	0.01	0.61	0.02	0.56	0.01
β_{ethn2}	−0.30	0.09	0.09	0.08	0.09	0.08	0.09	0.09	−0.31	0.10	−0.09	0.14
β_{ethn3}	−0.01	0.04	−0.18	0.04	−0.18	0.04	−0.19	0.05	−0.01	0.05	0.01	0.06
β_{educ2}	−0.92	0.04	−0.83	0.03	−0.83	0.03	−0.82	0.04	−0.91	0.04	−1.03	0.04
β_{educ3}	−1.19	0.04	−1.07	0.03	−1.07	0.04	−1.09	0.04	−1.18	0.04	−1.12	0.04
β_{sex}	−0.18	0.03	−0.01	0.01	−0.01	0.00	−0.17	0.03	−0.18	0.03	−0.15	0.03

[a] Age per 10 years.

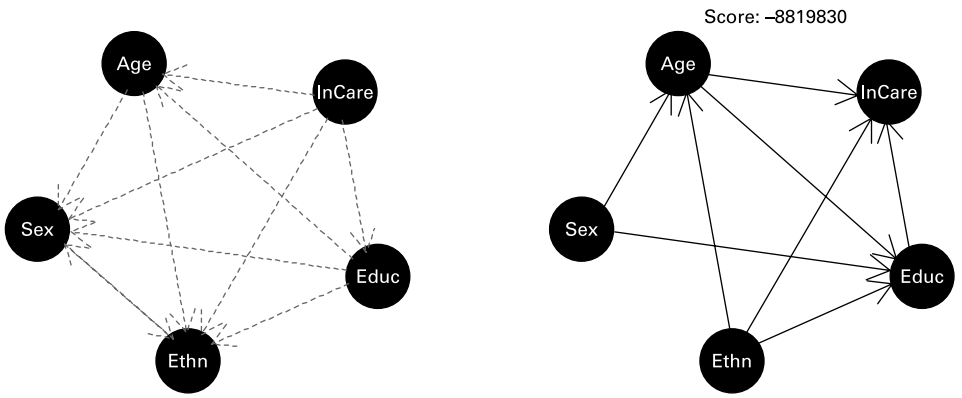

Fig. 2. Model B – its prior network (left), with banned arrows shown as dotted lines, and the result of a "greedy" search (right) [Nodes: InCare, institutional care; Educ, highest educational qualification; Ethn, ethnicity]

Model A is better at predicting the observed data than Model B. By ruling out an arrow between sex and ethnicity, Model B is simpler than Model A but not as good at predicting the observed data.

Inference from synthetic data generated by Models A and B suggests that institutional care is associated with all other variables except sex (Table 1: $\hat{\beta}_{sex} = -0.01$). Yet there is good support in the real data for an association between institutional care and sex – the point estimate is six times its standard error (Table 1: $\hat{\beta}_{sex} = -0.18$, $SE_{\hat{\beta}} = 0.03$). In Model C, we manually add an arrow between institutional care and sex to Model B (Figure 3 left). Data generated by Model C reproduce the association seen in the real data between institutional care and sex (Table 1: $\hat{\beta}_{sex} = -0.17$, $SE_{\hat{\beta}} = 0.03$). This is consistent with Reiter's observation that relationships not specified in an imputation model cannot be recovered from its synthetic data (Reiter 2005, p. 194).

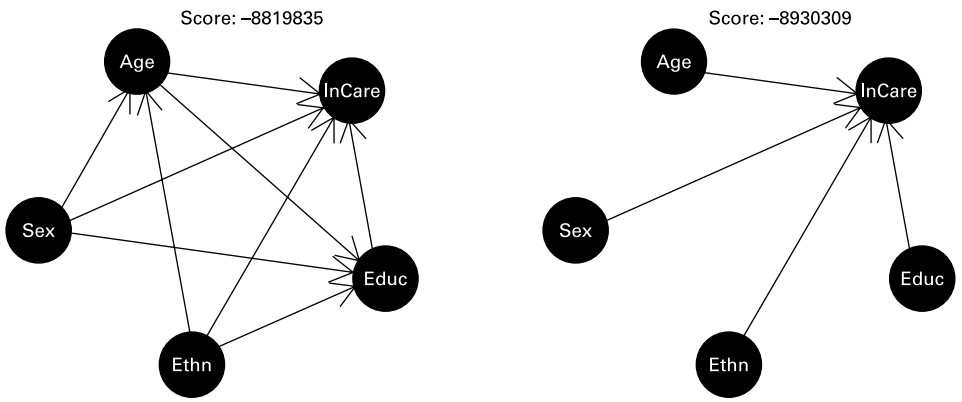

Fig. 3. Model C (left) and Model E (right) – Model D has the same network structure as C, but both Models D and E are fit with an imaginary database of size one, whereas Model C is fit with an imaginary database of size 720 (the default for these data) [Nodes: InCare, institutional care; Educ, highest educational qualification; Ethn, ethnicity]

Synthetic data from Models A to C give consistently poor estimates for ethnicity parameters. The data for this variable are predominantly in one category, and this leads us to suspect that prior information is having an unwarranted influence on inference. Our prior essentially adds two imaginary observations to the count in each cell. The influence of these prior observations will be greatest in cells with low counts and may affect parameter estimates associated with rare subgroups. In Model D, we fit Model C again but set the size of the imaginary database to one to largely remove the influence of the prior. Data generated by Model D reproduce all associations seen in the real data between institutional care and other variables (Table 1).

On the other hand, it is not enough to simply assert a network with arrows from all other variables to institutional care (Model E – Figure 3 right). Synthetic data from Model E give appreciably worse estimates for most parameters than data from Model D (Table 1). This parallels the relationship between logit and loglinear models. The loglinear model that corresponds to a given logit model has associations between the response (i.e., institutional care) and each logit predictor (i.e., each of the other variables) and is saturated with respect to associations between all logit predictors (see Agresti 1990, pp. 152–153). So to create synthetic data with good inferential properties under logistic regression, it is necessary to assert a network with arrows from all predictor variables to institutional care and with arrows between all (or nearly all) pairs of predictor variables.

Variability in real cell counts declines from Models B through to D for a given median synthetic cell count (Figure 4). As expected, synthetic data with better inferential properties have a higher risk of predictive disclosure (see Reiter 2005, pp. 188–189). In synthetic data from Model B, none of the 360 cells have a median count of one; in

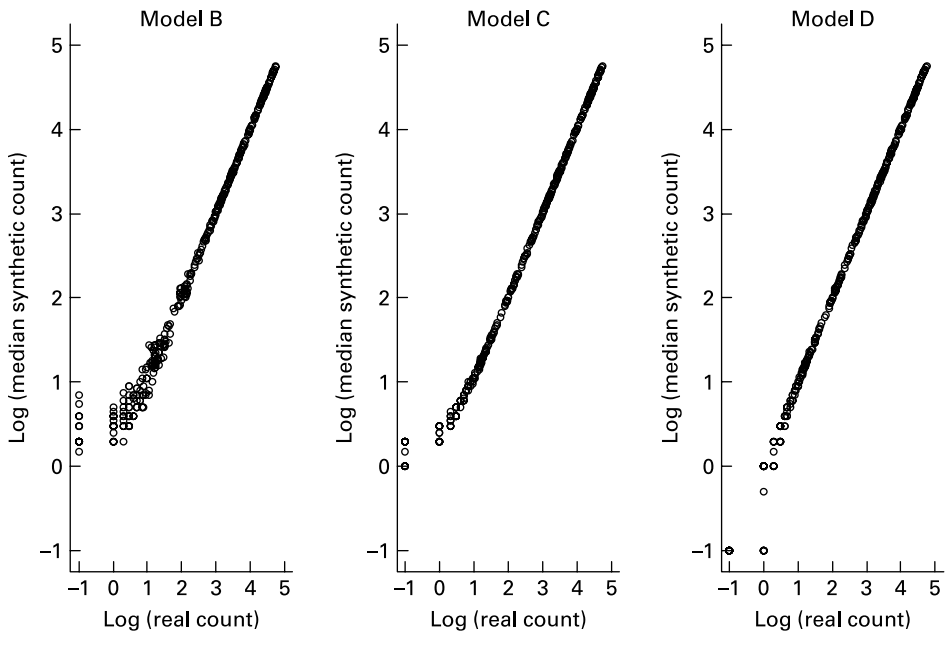

Fig. 4. Plots of median synthetic count versus real count in each cell for synthetic data from Models B, C, and D. Note that cell counts of zero have been shifted to 0.1 and therefore show as − 1 on each log scale

synthetic data from Model C, there are seven cells with a median count of one but none of these cells correspond to a unique individual in the real data. However, in synthetic data from Model D there are 20 cells with a median count of one and among these, 14 are cells with a unique individual in the real data. The influence of prior information is also obvious in these plots, as the size of the imaginary prior database inflates small cell counts in synthetic data from Models B and C.

5. Bayesian Model Averaging

5.1. Methods

The Bayesian equivalent to hypothesis testing is to choose the model with the highest posterior probability among competing models, where each model represents an alternative hypothesis. Assuming only two models are possible (H_1 or H_2), the posterior probability of the first model given data D is (Kass and Raftery 1995):

$$p(H_1|D) = \frac{p(D|H_1)p(H_1)}{p(D|H_1)p(H_1) + p(D|H_2)p(H_2)} \tag{1}$$

The ratio of posterior probabilities can be used to compare the two models so that:

$$\begin{aligned} &\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)} . \frac{p(H_1)}{p(H_2)}, \\ &\textit{ie posterior odds} = \textit{Bayes factor} \times \textit{prior odds} \end{aligned} \tag{2}$$

The probability of the data under one model relative to their probability under another is known as a Bayes factor. In "deal," the user cannot specify prior probabilities for different network models. Instead, at each step in a "greedy" search algorithm, the network model with the higher Bayes factor is chosen.

When a model has unknown parameters θ, calculating the probability of the data under that model requires integration over the parameter space of θ (Kass and Raftery 1995):

$$p(D|H_k) = \int p(D|\theta_k H_k)\pi(\theta_k|H_k)\partial\theta \tag{3}$$

In Equation 3, the first term within the integral is the likelihood function of θ_k under Model k and the second term is its prior distribution. Often numerical or approximate methods are required to evaluate this integral, but exact evaluation is possible for data from a distribution belonging to the exponential family of distributions if a conjugate prior distribution is used for θ (Kass and Raftery 1995). Hence the use in "deal" of the conjugate Dirichlet prior for multinomial probabilities. Heckerman et al. (1995) show that under certain assumptions, the probability of the data given a network (Equation 3) can be calculated for any network given a prior network and an imaginary database size as a measure of confidence in that prior network. These assumptions are parameter independence – the parameters associated with each node in a network are independent; parameter modularity – if a node has the same parents in another network, its parameters are the same in both networks; and likelihood equivalence – all networks that imply the same set of conditional probability distributions have the same likelihood. They suggest

that these assumptions should apply to complete observational data, but may not apply if missing data or experimental data cause a user to be more certain about some parts of a network than others (Heckerman et al. 1995, pp. 224–245). Cowell et al. (1999, pp. 261–263) discuss more flexible variants of the Dirichlet prior for expressing different degrees of certainty about different parts of a network.

For a network consisting entirely of discrete nodes, "deal" selects a "best" model according to the approach of Heckerman et al. (Bottcher and Dethlefsen 2003, pp. 7–9, 11–13). In a "greedy" search: (1) Equation 3 is evaluated for the prior network; (2) Equation 3 is evaluated for all other networks that differ by a single arrow from the prior network; (3) the network with the highest Bayes factor then replaces the prior network and Steps 2 and 3 are repeated until no network is found with a higher Bayes factor. When, however, there is uncertainty about the true model, using a single "best" model to create synthetic data may lead to inferences from that data that are unrealistically precise. The standard Bayesian solution to this problem is to average over a variety of models (Cowell et al. 1999, p. 250). Madigan and Raftery (1994) give examples where model averaging improves the predictive performance of graphical models.

When there are many (K) models to consider, the posterior probability of the kth model is:

$$p(H_k|D) = \frac{p(D|H_k)p(H_k)}{\sum_{k=1}^{K} p(D|H_k)p(H_k)} \tag{4}$$

The posterior distribution for some quantity of interest is "averaged" over many models by weighting the posterior distribution for each model by the posterior probability of that model where this is calculated using Equation 4 (Madigan and Raftery 1994). When creating multiple sets of synthetic data, this averaging can be achieved by repeatedly (1) selecting a model from among the set of all possible models using multinomial sampling where the probability of selecting each model is calculated using Equation 4, and (2) generating a single synthetic data set from the model selected. However, the problem is that as the number of nodes in the network increases, the increase in the number of possible network models is more than exponential (see Cowell et al. 1999, p. 256). Madigan and Raftery (1994) recommend averaging over a much smaller set of models, excluding models that are very unlikely compared to the most likely model and models that have more likely and simpler models nested within them. They use a variant of the "greedy" search algorithm to select a set of potentially acceptable models before applying these two exclusion criteria. We implement a version of their solution (Young 2008a, Section 8.3), applying only the first of these two exclusion criteria to the models found in a "greedy" search. There is some evidence that this simpler version has a slightly better predictive performance (Kass and Raftery 1995, pp. 146–147), as one would expect given that this version excludes fewer models from the set of all possible models.

5.2. Results

Starting with a prior network with no arrows, model averaging returns a set of acceptable models with just two members. Without a list of banned arrows, Model A is returned with posterior probability 0.993 and a variant of this model, with an arrow from sex to

institutional care, is returned with posterior probability 0.007. Adding a list of banned arrows leads to a similar result: Model B is returned with posterior probability 0.993 and Model C is returned with posterior probability 0.007. Obviously in both situations there is no point in creating synthetic data sets using Bayesian model averaging. Almost all data sets would be created under Models A or B respectively, and then inference from the resulting synthetic data would be no different from that already reported.

The problem is an example of "Bartlett's paradox" (Kass and Raftery 1995, p. 782; also known as the "Jeffreys – Lindley paradox" – Kass 1992, p. 555): conclusions based on Bayes factors appear to contradict conclusions based on estimation. Here model selection based on Bayes factors leads to models without arrows between sex and institutional care, yet there is good support in the real data for an association between the two. This is a consequence of using noninformative priors – a noninformative prior on a certain model parameter lends support to a model without that parameter (Kass and Raftery 1995, p. 782). When estimating a parameter, the influence of the prior rapidly diminishes with increasing sample size. When calculating a Bayes factor, the prior density is evaluated in each marginal likelihood (Equation 3). In large samples, the prior density can be thought of as an additive contribution to the log marginal likelihood (see Kass 1992, p. 555). While this contribution will eventually be overwhelmed by the contribution of the data, "even for large data-sets, the contribution of the prior can be significant" (Cowell et al. 1999, p. 249). Because the prior still contributes to the marginal likelihood even with a large sample and because a non-informative prior provides support for a simpler model, if a noninformative prior is used then a simple model carries more weight in model averaging than a more complex alternative.

To illustrate the influence of uninformative priors, we carry out model averaging for different sizes of imaginary databases. A larger imaginary database implies greater confidence in the equal conditional probabilities assigned to each node *a priori*. With a list of banned arrows, the default size is set to 720 and this leads to Model B. If the imaginary database size is set to 1, this leads to a single simpler model with no arrow from either sex or ethnicity to institutional care (Figure 5 left). If the imaginary database size is set to

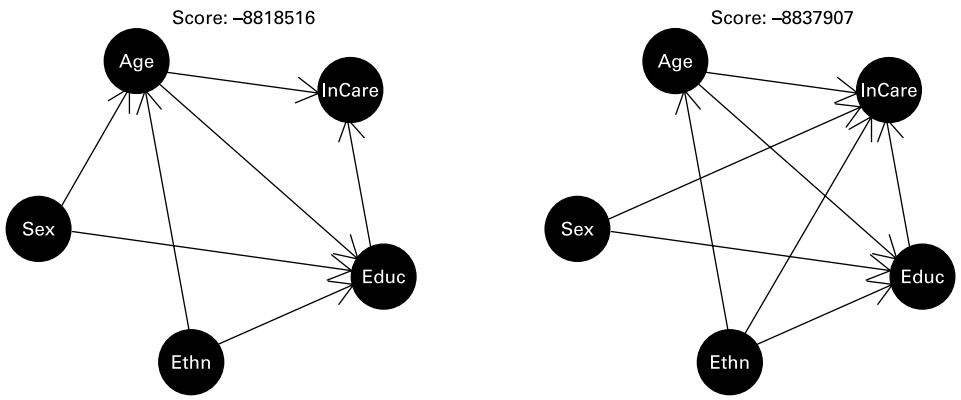

Fig. 5. Model averaging given the prior network of Model B leads to a single simple network (left) when the size of the imaginary database set to 1, and to a single more complex network (right) when the size of the imaginary database is set to 5,000 [Nodes: InCare, institutional care; Educ, highest educational qualification; Ethn, ethnicity]

5,000, this leads to a single more complex model with arrows from all other variables to institutional care (Figure 5 right).

This creates a dilemma: expressing a high level of confidence in an equal conditional probability prior leads to a more suitable model but as Table 1 shows, this prior will then have an unwanted influence on inference from synthetic data generated by this model. The solution is to assert informative priors for each node, rather than accepting the default prior of equal conditional probability for each category. Therefore we refit a prior network with no arrows, but with a list of banned arrows and assert prior probability distributions for each node. This is relatively simple to do for this prior network because there are no arrows between nodes. We consider that our assertions (Young 2008a, Section 8.4) represent knowledge equivalent to a database of 5,000 observations. These assertions lead to a set of two models (Figure 6). Neither model has an arrow from ethnicity to institutional care. The model with an arrow from sex to institutional care (Figure 6 right) is returned with posterior probability of 0.101. Results for synthetic data "averaged" over these two models will be given in next section.

All these examples lead to a set of acceptable models with at best two members and even then, with one model far more likely than the other. In these situations, inference from synthetic data will be much the same with or without Bayesian model averaging. Heckerman et al. (1995, p. 235) had the same experience: they investigated the effect of using more than one network structure to represent a joint distribution and were surprised by how little improvement this gave – "given a large data base, one network structure typically has a posterior probability far greater than the next most likely structure." Therefore without additional prior information, there is nothing to be gained from model averaging. Additional information could take the form of prior model probabilities. A number of strategies have been proposed based on the reasonable proposition that structures more closely resembling the prior network should have higher prior probabilities (Heckerman et al. 1995, p. 225; Madigan and Raftery 1994, p. 1544). Nevertheless, it seems unlikely that these strategies would, in the example above, lead to

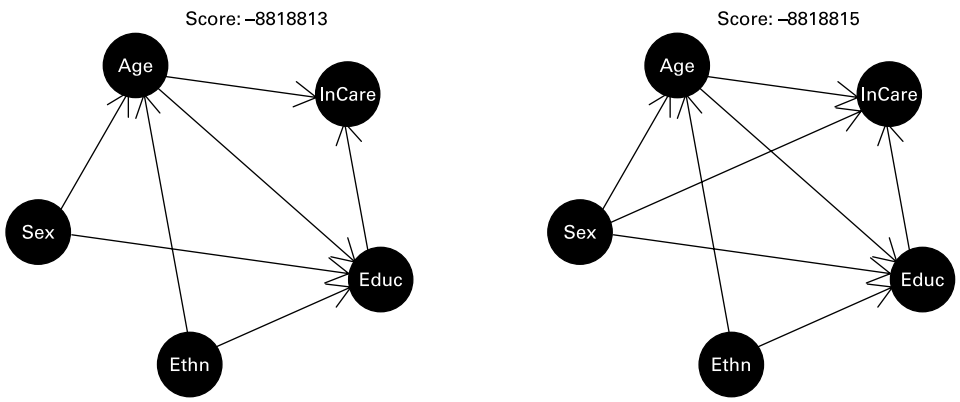

Fig. 6. Model averaging given the prior network of Model B, prior information for each node and an imaginary database of size 5,000. The most likely model (left) is returned with posterior probability 0.899; the other model (right) is returned with posterior probability 0.101 [Nodes: InCare, institutional care; Educ, highest educational qualification; Ethn, ethnicity]

much greater weight being placed on a model with a link between sex and institutional care given that this model differs from the most probable model by only one link and given the weight of evidence in favour of this most probable model.

The alternative is to provide a more realistic prior network, and to assert informative priors for its more complex conditional probabilities. For example, one might have to assert probabilities for being in institutional care for each combination of sex and ethnicity. Clearly this alternative would be unpalatable for an official statistics agency looking to automate the process of creating synthetic data.

6. Comparison With a Hierarchical Bayesian Imputation Model

6.1. Methods

Graham and Penny (2007) and Graham et al. (2009) take a hierarchical Bayesian approach to creating synthetic data for this institutional care data set. Their approach is to assert a prior Poisson log-linear model for the cross-classified data. The posterior mean count in each cell of the cross-classified data is then a weighted average of the count expected under this prior model and the observed cell count; the weight given to the prior model depends on how well it fits the data, with a poorly fitting prior model assigned less weight. Synthetic data from a hierarchical Bayesian model will, on average, replicate this posterior mean cell count, which depends on both the imputer's prior model and the observed data, whereas synthetic data from a conventional generalised linear model will, on average, replicate the model specified by the imputer. Hence synthetic data from a hierarchical Bayesian model are more robust to model misspecification than synthetic data from a conventional generalised linear model (Graham et al. 2009).

Here we compare the properties of synthetic data from three different imputation models. The three models are: (1) a hierarchical Bayesian model, (2) Bayesian network Model D, and (3) Bayesian model averaging over the two networks in Figure 6. The hierarchical Bayesian model is fit with prototype software (Young 2008b). As a prior model we specify a log-linear model with main effects, all two variable interactions, and all three variable interactions involving institutional care (i.e., the intermediate prior imputation model of Graham et al. 2009). With this prior we are specifying all the associations we would want the analyst to be able to recover from synthetic data through logistic regression with institutional care as the response. However, our prior does not give a saturated model for associations between the other explanatory variables, as in a logistic regression model. Using a more complex prior model that is saturated with respect to associations between explanatory variables might give synthetic data with a greater risk of disclosure and this relatively simple alternative has been shown to create synthetic data with satisfactory inferential properties (Graham et al. 2009).

6.2. Results

Inference from logistic regression is almost identical using either the real data, or synthetic data from the hierarchical Bayesian model, or synthetic data from Bayesian network Model D (Figure 3 left, fit with an imaginary database of size one). This applies to both the main effects logistic regression model (Table 2) and to the one way interaction model

Table 2. Parameter estimates and their standard errors for a main effects logistic regression model fit to the institutional care data and to synthetic data from: a hierarchical Bayesian model (HB), Bayesian network Model D, and Bayesian model averaging (BMA)

	Real data		HB Model		Model D		BMA	
	$\hat{\beta}$	$SE_{\hat{\beta}}$	$\hat{\beta}$	$SE_{\hat{\beta}}$	$\hat{\beta}$	$SE_{\hat{\beta}}$	$\hat{\beta}$	$SE_{\hat{\beta}}$
β_0	−8.15	0.07	−8.16	0.09	−8.16	0.11	−8.19	0.09
β_{age^a}	0.61	0.01	0.61	0.01	0.61	0.02	0.61	0.02
β_{ethn2}	−0.30	0.09	−0.28	0.09	−0.31	0.10	−0.01	0.05
β_{ethn3}	−0.01	0.04	−0.01	0.05	−0.01	0.05	−0.09	0.01
β_{educ2}	−0.92	0.04	−0.91	0.03	−0.91	0.04	−0.91	0.04
β_{educ3}	−1.19	0.04	−1.19	0.04	−1.18	0.04	−1.16	0.04
β_{sex}	−0.18	0.03	−0.18	0.03	−0.18	0.03	−0.03	0.06

[a] Age per 10 years.

(Table 3). Using synthetic data from Bayesian model averaging, inference is poor when it involves either sex or ethnicity parameters, reflecting the absence of arrows in the most likely network from these two nodes to institutional care (Figure 6 left).

Although the hierarchical Bayesian model and Model D provide synthetic data with similar inferential properties, real cell counts are more variable for a given median

Table 3. Parameter estimates and their standard errors for a one way interaction logistic regression model fit to the institutional care data and to synthetic data from: a hierarchical Bayesian model (HB), Bayesian network model D, and Bayesian model averaging (BMA)

	Real data		HB Model		Model D		BMA	
	$\hat{\beta}$	$SE_{\hat{\beta}}$	$\hat{\beta}$	$SE_{\hat{\beta}}$	$\hat{\beta}$	$SE_{\hat{\beta}}$	$\hat{\beta}$	$SE_{\hat{\beta}}$
β_0	−6.49	0.16	−6.52	0.17	−6.49	0.19	−6.61	0.18
β_{age^a}	0.27	0.03	0.28	0.04	0.27	0.05	0.29	0.04
β_{ethn2}	−0.52	0.34	−0.57	0.49	−0.62	0.54	−0.05	0.13
β_{ethn3}	−0.32	0.16	−0.30	0.19	−0.33	0.19	−0.91	0.10
β_{educ2}	−2.55	0.19	−2.52	0.22	−2.53	0.24	−2.80	0.16
β_{educ3}	−3.69	0.20	−3.68	0.22	−3.69	0.22	−3.86	0.20
β_{sex}	−1.28	0.13	−1.29	0.18	−1.30	0.16	−0.20	0.32
$\beta_{age^a \times ethn2}$	−0.02	0.07	−0.01	0.10	0.00	0.13	0.01	0.05
$\beta_{age^a \times ethn3}$	0.10	0.03	0.10	0.04	0.11	0.04	0.18	0.03
$\beta_{age^a \times educ2}$	0.40	0.03	0.39	0.04	0.40	0.04	0.34	0.04
$\beta_{age^a \times educ3}$	0.54	0.03	0.54	0.04	0.54	0.04	0.49	0.04
$\beta_{age^a \times sex}$	0.18	0.02	0.18	0.03	0.19	0.03	0.04	0.05
$\beta_{ethn2 \times educ2}$	0.48	0.26	0.50	0.32	0.40	0.32	−0.02	0.28
$\beta_{ethn3 \times educ2}$	−0.59	0.13	−0.61	0.10	−0.61	0.14	0.03	0.06
$\beta_{ethn2 \times educ3}$	0.89	0.24	0.91	0.24	0.87	0.21	0.04	0.11
$\beta_{ethn3 \times educ3}$	−0.74	0.13	−0.74	0.16	−0.71	0.16	0.00	0.16
$\beta_{ethn2 \times sex}$	0.31	0.18	0.31	0.16	0.30	0.18	0.02	0.16
$\beta_{ethn3 \times sex}$	0.05	0.09	0.07	0.08	0.05	0.09	−0.05	0.09
$\beta_{educ2 \times sex}$	−0.10	0.07	−0.10	0.07	−0.10	0.07	0.01	0.06
$\beta_{educ3 \times sex}$	0.36	0.07	0.38	0.07	0.36	0.08	0.04	0.10

[a] Age per 10 years.

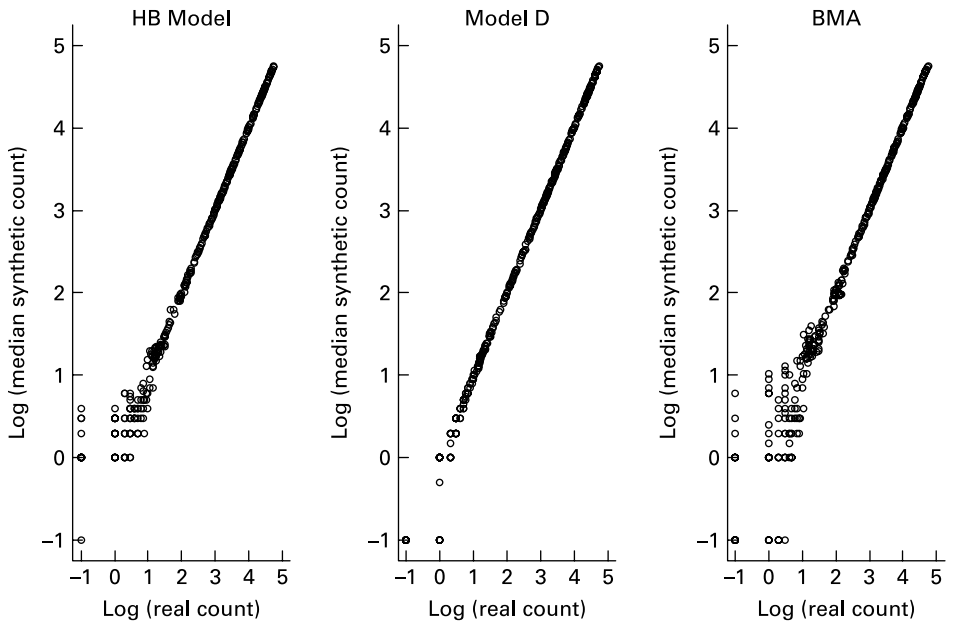

Fig. 7. Plots of median synthetic count versus real count in each cell for synthetic data from: a hierarchical Bayesian (HB) model, Bayesian network Model D and Bayesian model averaging (BMA). Note that cell counts of zero have been shifted to 0.1 and therefore show as −1 on each log scale

synthetic cell count in synthetic data from the hierarchical Bayesian model (Figure 7). This implies that the predictive disclosure risk is lower in synthetic data from the hierarchical Bayesian model than in synthetic data from Model D. For example, there are 28 cells with a median count of one in synthetic data from the hierarchical Bayesian model and 8 (29%) of these cells correspond to a unique individual in the real data. In synthetic data from Model D, there are 20 cells with a median count of one, and 14 (70%) of these cells correspond to a unique individual in the real data.

7. Discussion

These are somewhat disappointing results from the perspective of an official statistics agency looking to use a Bayesian network to automate the process of creating synthetic data. It would seem relatively easy to assert a single Bayesian network that will create synthetic data with the same inferential properties as the real data. First specify a natural order among variables via a list of banned arrows. Second assert a network with arrows between all nodes other than arrows that are banned. Third specify a very small imaginary database and fit the network to the data. But such a network would create synthetic data that are unnecessarily precise with a higher risk of disclosure than is probably acceptable. A hierarchical Bayesian model could create synthetic data that are robust to model misspecification and have a lower risk of disclosure.

In theory Bayesian model averaging should solve this problem. Synthetic data would then be created from a mixture of likely networks so that the data are more variable and robust to network misspecification. But with little prior information, the most likely

networks tend to be too simplistic and these lead to synthetic data with poor inferential properties. Prior information in the form of prior network probabilities may not be enough to steer the selection process from a simple prior network towards more complex alternatives. It would seem necessary to start the selection process by specifying a more realistic prior network to begin with and to specify informative prior probabilities for its nodes (otherwise simple models carry more weight in model averaging than their more complex alternatives). Such a selection process would not be easy to automate.

Others have their reservations about using Bayes factors for model selection and for model averaging. Commenting on the use of Bayes factors for these purposes, Gelman and Rubin (1995, p. 170) note that if the goal is to produce accurate parameter estimates rather than select a parsimonious model, then "a hierarchical model might be more compelling." These results support their position. A key difference between Bayesian model averaging and a hierarchical Bayesian model is in the influence of prior information. This information is critical for assigning appropriate weights to candidate networks in model averaging if synthetic data are to have good inferential properties and an acceptable risk of disclosure. In contrast, simulation results in Graham et al. (2009) show that the hierarchical Bayesian model is relatively robust to the choice of a prior model – synthetic data can have acceptable inferential properties even if this prior model is not as complex as the model subsequently used to analyse the synthetic data.

Bayesian networks may be more successful as imputation models in scenarios other than the one we consider here. Bayesian networks have been successfully applied to data sets comprising many hundreds of variables (Tsamardinos et al. 2006). Official statistical agencies sometimes release unit record data sets with many variables, such as the Public Use Microdata Sample file from the U.S. Census Bureau, which contains 67 variables collected in the American Community Survey (http://www.census.gov/acs/www/Products/PUMS/index.htm). With high-dimensional data it would be difficult if not impossible to create synthetic data using generalised linear or hierarchical Bayesian imputation models. If an agency were willing to put effort into asserting an informative structure and probabilities for a set of variables (rather than just run some automatic or semiautomatic process), then synthetic data from Bayesian model averaging over a set of Bayesian networks might have sufficient analytic value for release as a general purpose approximation to that set of variables. This would allow analysts to develop hypotheses before making a formal request for synthetic data with a specific analytic focus with which to test those hypotheses.

Bayesian networks may also have a place in official statistics as a communication tool, rather than as an imputation model for creating synthetic data. Networks could be made available via the Internet so that users outside the official statistics agency can query a network about how likely a certain value of one variable is given the values of other variables (see Ramoni and Sebastiani 2001, pp. 434–436). This is the conventional use of a Bayesian network and commercial software is available for querying a network. With these tools, a user can make descriptive inferences from data but not analytic inferences; synthetic data can be used for both (see Reiter 2005, pp. 194–197) but a network offers the advantages of speed and simplicity.

8. References

Agresti, A. (1990). Categorical Data Analysis. New York: John Wiley.

Albert, J.H. (1988). Computational Methods Using a Bayesian Hierarchical Generalized Linear Model. Journal of the American Statistical Association, 83, 1037–1044.

Bottcher, S.G. and Dethlefsen, C. (2003). deal: A Package for Learning Bayesian Networks. Journal of Statistical Software, 8, 1–40.

Christiansen, C.L. and Morris, C.N. (1997). Hierarchical Poisson Regression Modeling. Journal of the American Statistical Association, 92, 618–632.

Cowell, R.G., Dawid, A.P., Lauritzen, S.L., and Spiegelhalter, D.J. (1999). Probabilistic Networks and Expert Systems. New York: Springer-Verlag.

Di Zio, M., Scanu, M., Coppola, L., Luzi, O., and Ponti, A. (2004). Bayesian Networks for Imputation. Journal of the Royal Statistical Society, Series A, 167, 309–322.

Fienberg, S.E. and Makov, U.E. (1998). Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data. Journal of Official Statistics, 14, 385–397.

Gelman, A. and Rubin, D.B. (1995). Avoiding Model Selection in Bayesian Social Research. Sociological Methodology, 25, 165–173.

Graham, P. and Penny, R. (2007). Multiply Imputed Synthetic Data Files. Official Statistics Research Series Volume 1, Wellington, New Zealand: Statistics New Zealand. http://www.statisphere.govt.nz/official-statistics-research/series/volume-1-2007

Graham, P., Young, J., and Penny, R. (2009). Multiply Imputed Synthetic Data: Evaluation of Hierarchical Bayesian Imputation Models. Journal of Official Statistics, 25, 245–268.

Heckerman, D., Geiger, D., and Chickering, D.M. (1995). Learning Bayesian Networks: the Combination of Knowledge and Statistical Data. Machine Learning, 20, 197–243.

Jensen, F.V. (2001). Bayesian Networks and Decision Graphs. New York: Springer-Verlag.

Kass, R.E. (1992). Bayes Factors in Practice. The Statistician, 42, 551–560.

Kass, R.E. and Raftery, A.E. (1995). Bayes Factors. Journal of the American Statistical Association, 90, 773–795.

Madigan, D. and Raftery, A.E. (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. Journal of the American Statistical Association, 89, 1535–1546.

Neapolitan, R.E. (2004). Learning Bayesian Networks. Upper Saddle River, New Jersey: Pearson Prentice Hall.

R Development Core Team (2004). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. http://www.R-project.org

Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, 19, 1–16.

Ramoni, M. and Sebastiani, P. (2001). Analysis of Survey Data With Bayesian Networks. Bayesian Methods With Applications to Science, Policy, and Official Statistics: Selected Papers From the Sixth World Meeting of the International Society for Bayesian Analysis (ISBA 2000), E.I. George (ed.). Luxembourg: Eurostat. http://www.stat.cmu.edu/ISBA

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. Journal of the Royal Statistical Society, Series A, 168, 185–205.

Rubin, D.B. (1993). Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, 9, 461–468.

Russell, S.J. and Norvig, P. (1995). Artificial Intelligence: A Modern Approach. Upper Saddle River, New Jersey: Prentice Hall.

Thibaudeau, Y. and Winkler, W.E. (2002). Bayesian Networks Representations, Generalized Imputation, and Synthetic Micro-Data Satisfying Analytic Constraints. Statistical Research Report Series. Washington, DC: U.S. Bureau of the Census. http://www.census.gov/srd/www/byyear.html

Tsamardinos, I., Brown, L.E., and Aliferis, C.F. (2006). The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. Machine Learning, 65, 31–78.

Young, J. (2008a). Using Bayesian Networks to Create Synthetic Data. In Methods for Creating Synthetic Data, eds. P. Graham, J. Young, and R. Penny, Official Statistics Research Series Volume 3, Wellington, New Zealand: Statistics New Zealand. http://www.statisphere.govt.nz/official-statistics-research/series/volume-3-2008.

Young, J. (2008b). Synthetic Tables: A Prototype for Creating Synthetic Census Tables. In Methods for Creating Synthetic Data, eds. P. Graham, J. Young, and R. Penny, Official Statistics Research Series Volume 3, Wellington, New Zealand: Statistics New Zealand. http://www.statisphere.govt.nz/official-statistics-research/series/volume-3-2008

Received May 2007
Revised February 2009