A Technique for Incorporating Data Missing Not at Random (MNAR) into Bayesian Networks

VALERIE SESSIONS, Charleston Southern University
JUSTIN GRIEVES, Charleston Southern University
STANLEY PERRINE, Georgia Gwinnett College

Data Quality, Missing Not at Random, Bayesian Networks

We present a technique for incorporating data attributes that are supposed Missing Not at Random (MNAR) into Bayesian Networks (BNs). While traditional methods of incorporating data that is Missing at Random (MAR) into BNs are well documented, there are fewer tested methods for discovering and incorporating data Missing Not at Random (MNAR). We present a review of literature in BNs and missing data, an illustrative example of our method, test setup and results, as well as limitations and future research avenues. It is our eventual goal to develop from this technique a method to discover whether the missing mechanism is Missing at Random (MAR) or Missing Not at Random (MNAR).

1. INTRODUCTION

Managing missing data is an active topic of research in the fields of data quality, computer science, and statistical analysis. We shall adopt the common categories of missing data found in [Little and Rubin 1987] of Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR) in this paper. We have several methods of handling data that is categorized MCAR or MAR – multiple imputation, listwise deletion and marginalization among others. There are fewer methods of handling data classified MNAR; in fact, there are few methods that can successfully identify missing data as MNAR. Many of these methods are reviewed in [Tremblay et al 2010]. Tremblay et. al. developed a method of detecting and categorizing MNAR data using Association Rule Mining which may prove to be very useful in determining how to deal effectively with missing data. It is the authors' viewpoint that more research in this area is needed, as a great wealth of information may be waiting to be uncovered by simply looking more closely at why fields are empty. It could be due to such technical issues as poor calibration of an instrument or data points outside a sensor's range; it could be human error in the actual collection of the data, or possibly due to poor data entry standards. Of course, there are also possible physiological motivations, such as embarrassment over an answer on a questionnaire or fear that an anonymous survey is not truly "anonymous".

Regardless of the reason(s) as to why a particular data set has items missing, the authors agree with Lin and Huag that often the 'missing-ness' of the data does indeed have its own story to tell [Lin and Haug 2008]; therefore, further research into both the proper labeling of missing data as MNAR as well as the proper handling of such MNAR data is vital to the continued improvement of data quality, as well as the rapidly increasing number of data-driven decisions being made using MNAR data.
The authors therefore propose a method for accommodating MNAR data in Bayesian Networks (BNs); this method may also serve to discover the missing mechanism.
The structure of this paper is as follows. First, we review the literature pertaining to data quality, statistical analysis and BNs. Second, we discuss our methodology and give a demonstration of our technique. Third, we discuss our rest setup and results. Finally, we present our conclusions regarding these results and present avenues for future research.

2. LITERATURE REVIEW

Data and information quality is a growing field of research that seeks to classify types of data error, map data flows or Information Products (IP) and create better business processes to minimize the risk of poor quality data, among other topics. For an overview of this large field the authors recommend [Lee et al 2006]. For our purposes here we will review the classifications of missing data and point to research that uses these methods or seeks to classify data as MAR or MNAR. Missing data, as classified by Little and Rubin [1987], can be MCAR, MAR, or MNAR. MCAR represents the case of a complete fluke – the missing mechanism is neither a result of the variable itself nor any other value in the distribution. MAR data can be missing because of its relationship to a particular variable in the model, but the missing mechanism has no relationship to other variables in the model. There is much research on how to account for both MCAR and MAR data. A good overview of these methods can be found in both [Horton and Klienman 2007] and [McKnight et al 2007]. As these methods are quire similar, we will in the remainder of this paper refer to this category only as MAR data. Methods for handling MAR data, however, are inappropriate for data MNAR and have been shown to be unsuccessful when used on MNAR data [Almedar 2009]. Research conducted by Tremblay, et. al. seeks to discover bias patterns in missing data using Association Rule Mining (ARM) algorithms, and has done so with some success. In addition, biomedical informatics research by [Lin and Haug 2008] incorporated missing data as an explicit classifier in BN modeling and found that in most cases the BN trained with the missing classifier performed better than those without.

[Ramoni and Sebastiani 1999] make significant progress towards an estimation method that is appropriate under cases of both MAR and MNAR (which they refer to as *Ignorable* and *Non Ignorable*) data with the Bound and Collapse (BC) algorithm which estimates the probability distribution. They compare this method to the EM and Gibb's Sampling (GS) and while the non ignorable estimation and prediction errors do not vary that much among the methods, the execution times for BC are significantly faster than EM or GS. Our method differs in that it does not seek to estimate the missing data but instead retains the fact that the data is missing.

Bayesian networks (BNs) are used to model a domain of knowledge. A BN consists of a set of variables and directed edges between these variables. Together these form a directed acyclic graph (DAG). To each variable A, there is a set of discrete states $a_1,...,a_n$. If A is a variable with states $a_1,...,a_n$, then $P(A)$ denotes a probability distribution over these states

$$P(A) = (x_1,...,x_n); \qquad x_i \geq 0; \qquad \sum_{i=0}^{n} x_i = 1$$

Also, if there exists a variable A with parents $B_1,...,B_n$, there is attached a potential table $P(A \mid B_1,...,B_n)$. For those less familiar with BNs, we use here an example, Wet Lawn, to illustrate these networks in a simplified manner. For a more detailed mathematical approach, we recommend [Cowell et al 1999], [Jensen 2001], [Jensen and Nielson 2007], and [Neapolitan 2004].

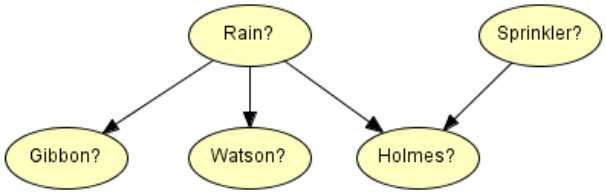


Figure 1. Wet Lawn

Mr. Holmes deduces whether his lawn is wet because it has rained or because his sprinkler is working. He gathers evidence by looking to see if Mr. Watson's and Mrs. Gibbon's lawns are also wet. The strength of these dependencies is modeled as a probability – typically represented by a conditional probability table. A graphical representation of the BN is shown above in Figure 1. The conditional probability tables for three nodes in our Wet Lawn example are shown in Figure 2.

Sprinkler?(Sprinkler)

yes	0.1
no	0.9

Rain?(Rain)

yes	0.1
no	0.9

Holmes?(Holmes)

Sprinkler	yes		no	
Rain	yes	no	yes	no
yes	1.0	0.9	0.99	0.0
no	0.0	0.1	0.01	1.0

Figure 2. Wet Lawn Probability Tables

As we see from this example, Holmes' wet lawn is dependent upon either the sprinkler or the rain. If it has not rained and if the sprinkler has not been working, the lawn will not be wet. If it has either rained or the sprinkler has been working, but not both, there is a 90-99% chance that it is wet; if it has both rained and the sprinkler is working, there is 100% chance that the lawn is wet.

Using Bayes Rule, or the Chain Rule, we can also update our beliefs about the network based on new evidence – either *Sprinkler? = "yes"* or *"No"* or *Rain? = "Yes" or "No"*. Bayes rule is stated as:

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

where $P(A \mid B)$ is the probability of A given B, $P(A)$ and $P(A)$ are the prior probabilities of A and B respectively, and $P(B \mid A)$ is the probability of B given A. Based on a combination of evidence and prior probability of a variable, the probability of certain results will increase or decrease (assuming the variables are dependent).

3. PROPOSED TECHNIQUE

Because the vast majority of previously attempted methods to adjust for or otherwise contend with MNAR data have seemed to fall short of desired results, we propose to

actually incorporate the missing "data point" into the BN – i.e. as referenced above, the missing-ness of the data is itself an integral piece of data.
Our technique is straightforward, and as we show below, can be applied and whether the data is MAR or MNAR. We mark missing data *'nr'* and treat this as an additional attribute for the node in question. For example, if a node's attributes are *'Yes'* and *'No'*, then with the missing attribute we now have '*Yes'*, '*No'* and *'nr'*. We then learn the structure and parameters with the missing data attribute and are able to form correlations based on this missing attribute. As one would assume, we will show that if the missing mechanism is added to a MAR data set, there is no apparent correlation between the *'nr'* attribute and other nodes. More importantly, however, we also show that adding the missing mechanism to MNAR data can lead to a correlation between the *'nr'* attribute and other nodes, thereby creating a better learned structure than simply deleting nodes.
The process is to augment the original probability distribution with an additional "missing" state – given a variable A, which has missing data, we add a state a_{NR} to the given states $a_1,...,a_n$ of A, with x_{NR} denoting the probability of this state occurring. Restating this definition above, we now have:

If A is a variable with states $a_1,...,a_n$ along with missing data state a_{NR}, then $P(A)$ denotes a probability distribution over these states

$$P(A)=(x_1,...,x_n,x_{NR}); \qquad x_i \geq 0; \qquad \sum_{i=0}^{n} x_i + x_{NR} = 1$$

There is no other change to the variables or their directed edges. A demonstration of this method shall be shown in the Illustrative Example section of this paper, but let us first give a little more information regarding our rationale for and confidence in choosing this method.

Definition: A data set is considered **complete** if each variable has no missing values; that is to say $x_{NR}=0$.

First note that adding the *'nr'* state and associated probability x_{NR} does not alter a data set that is complete. While the proof of this fact is trivial, we believe it important because we seek to present a method that can be used in all situations of missing data – including no missing data at all.

Lemma: Let $A=(a_1,a_2,...,a_n)$ be a variable in a complete data set, and let a_{NR} represent the state in which a data value is missing or otherwise unreported. Then, the variable $A'=A\cup\{a_{NR}\}$ is an extension of the variable A such that the corresponding distribution $P(A')$ restricted to A is equivalent to the original distribution: i.e. $P(A'=a_i)=P(A=a_i)$ for all $i=1,...,n$.

Proof: Assume $A=(a_1,a_2,...,a_n)$ is a variable in a complete data set, and define $A'=(a_1,a_2,...,a_n,a_{NR})$. Next, define the distribution $P(A')$ by $P(A'=a_i)=x_i$ for all $i=1,...,n$ and $P(A'=a_{NR})=0$
Since

$$0 \le x_i \le 1 \text{ for all } i=1,...,n;\ 0 \le x_{NR} \le 1$$

and

$$\sum_{a_j \in A'} P(A'=a_j) = \sum_{i=1}^{n} x_i + x_{NR} = \sum_{i=1}^{n} x_i + 0 = 1 \text{ (by definition of complete),}$$

$P(A')$ forms a probability distribution for A' with the desired equivalence.

Essentially, A' is a variable with the *"nr"* state "added in" without changing the probability for any of the nodes in A.

4. ILLUSTRATIVE EXAMPLE

In our example below, we assume the BN nodes, correlations, and prior probabilities we present have either been constructed by a Subject Matter Expert (SME) or through a structure learning (SL) algorithm. In our example, we will use the pre-created BN to then update probabilities based on new evidence, thus showing how such a BN would retain the nature of the missing attribute(s). In the next section of the paper, we will go farther and learn the structure of the BN from the data itself.

We demonstrate the usefulness of our method with the canonical BN Visit to Asia or Chest Clinic. Visit to Asia is a fictitious BN created by [Lauritzen and Spielgelhalter 1988] and shown in Figure 2.

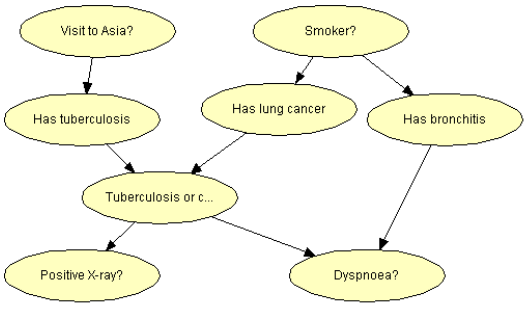


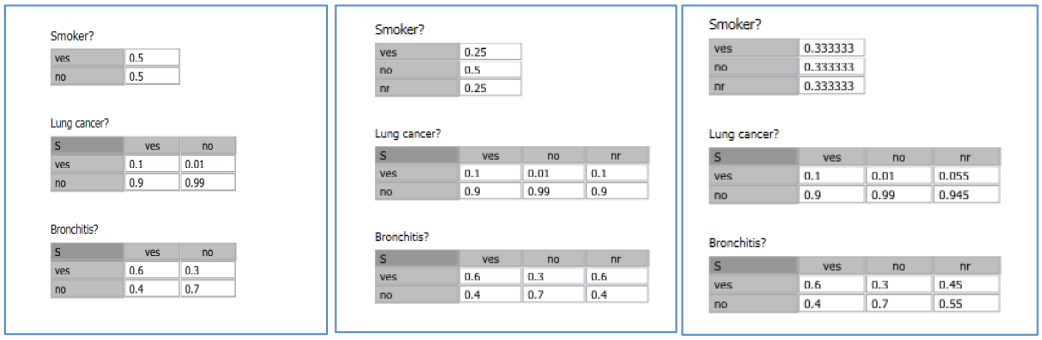
Figure 2: Visit to Asia

The network represents a situation as follows:

> Shortness-of-breath (dyspnoea) may be due to tuberculosis, lung cancer or bronchitis, or none of them, or more than one of them. A recent visit to Asia increases the chances of tuberculosis, while smoking is known to be a risk

> factor for both lung cancer and bronchitis. The results of a single chest X-ray do not discriminate between lung cancer and tuberculosis, as neither does the presence or absence of dyspnoea.
> [Lauritzen and Spielgelhalter 1988].

To demonstrate this method, we have added state *'nr'* to the variable "*Smoker?"*. In one instance we do so in a in a fashion that mimics MAR mechanism and another in a fashion that mimics the MNAR mechanism of missing-ness. We show sample prior probability tables for the original (perfect), MAR, and MNAR data in Figure 3. The left column is original data, the middle contains the missing data attribute *'nr'* in *a* MAR pattern of missing-ness, and the right column contains the missing data attribute *'nr'* in a MNAR pattern of missing-ness. To mimic the MAR mechanism of missing-ness, we have evenly split the prior probability of missing data between what would perceivably have been *Smoker? = "Yes" or Smoker? = "No"* responses. To mimic the MNAR mechanism, the prior probability is exclusively in the *Smoker = "Yes"* response, in other words *Smoker?="nr"* percentage of missing data is exclusively taken from the *Smoker?=""Yes"* category.



ORIGINAL **MNAR** **MAR**

Figure 3: Prior Probabilities Visit to Asia

Based on this BN and the prior probabilities presented in Figure 3, when evidence is presented in the form *Smoking? = nr*[1], then via Bayes Rules we arrive at the posterior probabilities presented in Table 4. Note that in the case of MAR data the chances of lung cancer have not changed from the original 'perfect' data's prior probability. However, with the MNAR mechanism, there is a larger correlation between *'nr'* and lung cancer. This is encouraging because it demonstrates that this method retains the original relationships in the data. If one were to simply delete all missing data rows, or use MAR methods such as MI, we would miss the correlation between *Smoker? = 'nr'* and a higher chance of lung cancer in the case of no response.

[1] We used the Hugin Decision Engine [Madsen et al 2005] to generate the posterior probabilities in our illustrative example.

	Posterior Probability with evidence *Smoking? = nr* (Difference from Prior)			
	Lung Cancer?		Bronchitis?	
	Yes	No	Yes	No
MNAR	**10 (4.5)**	**90 (-4.5)**	**60 (15)**	**40 (-15)**
MAR	**5.5 (0)**	**94.5 (0)**	**45 (0)**	**55 (0)**

Figure 4: Posterior Probabilities given *Smoker?='nr'*

5. TEST METHOD AND RESULTS

In our illustrative example, we used a BN whose prior probabilities had been set by either a SME or a SL algorithm in order to show the usefulness of retaining the 'missing-ness' attribute. In order to determine if this technique can also be used to determine the missing-ness mechanism – MAR or MNAR – we used a SL algorithm, 'Tabu', to learn our BN with various amounts of missing data – both with a MAR and MNAR mechanism. We implemented this via an R Code package called BNLearn [Scutari 2014] and two different data sets – Trip to Asia/Chest Clinic and Alarm - to test our technique. The Trip to Asia data set was explained above. Alarm [Beinlich et al 1989] is a data set that was developed to provide an alarm message system for patient monitoring. It has 37 variables relating to body monitoring elements such as blood pressure, lung ventilation, and heart rate.

Using a 5000 record set for Visit to Asia and a 20,000 record set for Alarm we degraded the data in the following manner.

Visit To Asia

1. Visit to Asia MAR – degraded the *Smoker?* node by replacing *Smoker='yes'* and *Smoker='no'* at random with the *'nr'* attribute at 1%, 3%, 5%, 10%, 20%, 30%, 40%, and 50% percentage of the data.

2. Visit to Asia MNAR – degraded the *Smoker?* node by replacing *Smoker = 'Yes'* with the *'nr'* attribute at 1%, 3%, 5%, 10%, 20%, 30%, 40%, and 50% percentage of the data.

3. Visit to Asia MAR Listwise delete – replicated the listwise deletion method by deleting all rows with the *'nr'* attribute from #1.

4. Visit to Asia MNAR Listwise delete - replicated the listwise deletion method by deleting all rows with the '*nr'* attribute from #2.

Alarm

1. Alarm MAR – degraded the *VLNG* node by replacing *VLNG='LOW'* and *VLNG = 'MEDIUM'* and *VLNG = 'HIGH'* at random with the *'nr'* attribute at 1%, 3%, 5%, 10%, 20%, 30%, 40%, and 50% percentage of the data.

2. Alarm MNAR – degraded the *VLNG* node by replacing *VLNG = 'LOW'* with the *'nr'* attribute at 1%, 3%, 5%, 10%, 20%, 30%, 40%, and 50% percentage of the data.

3. Alarm MAR Listwise delete – replicated the listwise deletion method by deleting all rows with the *'nr'* attribute from #1.

4. Alarm MNAR Listwise delete - replicated the listwise deletion method by deleting all rows with the *'nr'* attribute from #2.

We then used the 'Tabu' learning algorithm as implemented by the BNLearn package [Scutari 2014] to learn the BN for each percentage of missing data[2]. Then by cross validating the learned model with its data set we arrived at the log likelihood loss for each learned BN. The log likelihood loss shows the goodness of fit of the learned model – how closely the learned BN matches its data set. Our results for Visit to Asia are shown in Figure 5 and for Alarm in Figure 6, with the "gold standard" log likelihood loss added as a black line for clearer illustration. A deviation between the learned BN's log likelihood loss and the gold standard's represents a deviation from the BN that would be learned from the original data set (or complete data).

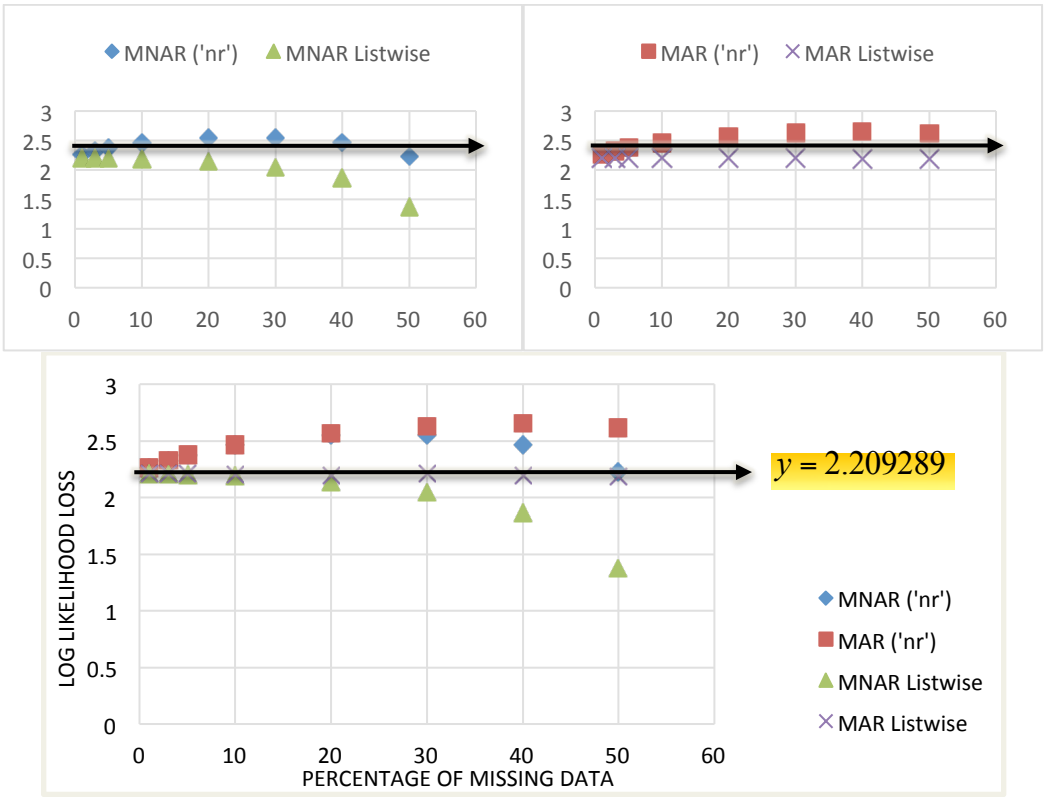


Figure 5: Comparison of methods on "Visit to Asia" data set

[2] The 'Tabu' method was chosen from the available learning algorithms by testing all methods on a 'gold standard' or complete data set. We then cross-validated each BN and chose the method resulting in the lowest log likelihood loss. As both the SL algorithm and the loss function are independent variables, other methods and loss functions could have been chosen, provided they remained constant throughout the testing.

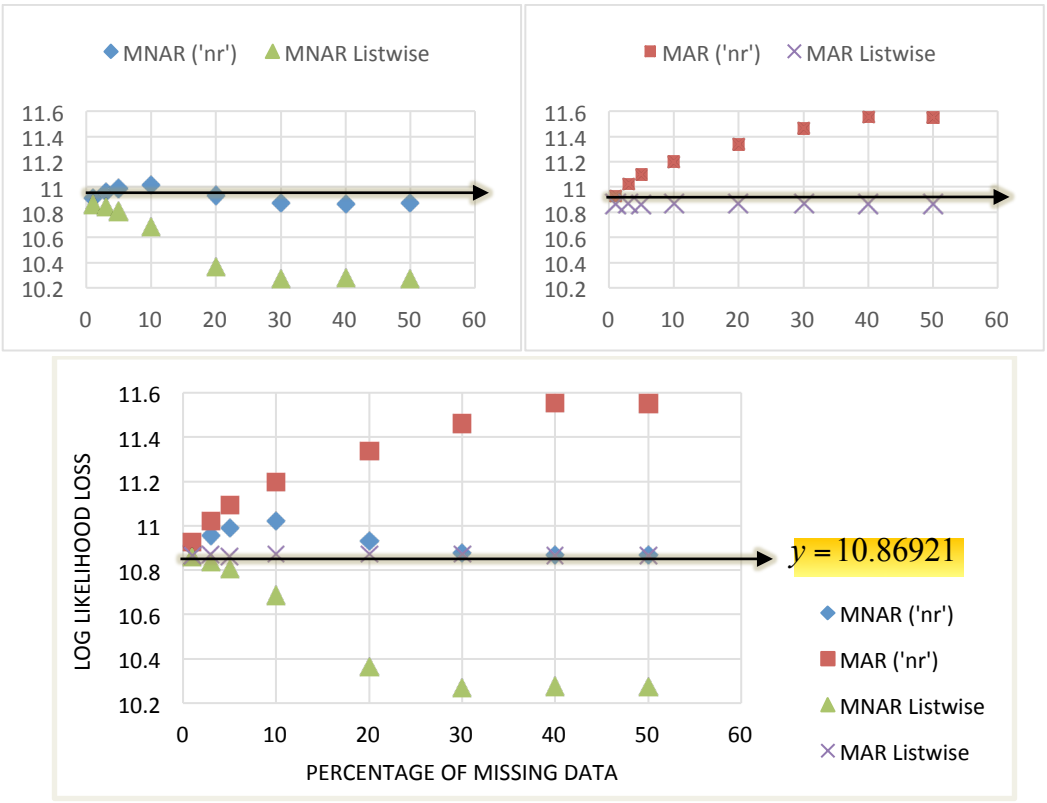


Figure 6: Comparison of methods on "Alarm" data set

6. DISCUSSION OF RESULTS

Observing the results for MAR data sets, we conclude that traditional listwise deletion method works well. The learned BNs created after listwise deletion was used on the data sets better conform to the "gold standard" learned BN. Adding the *'nr'* attribute and retaining the missing data degrades the model in a reasonable way – as the percentage of missing data increases, the farther the learned BN gets from the "gold standard" learned BN. However, for MNAR data, we have demonstrated that our technique affords a significant improvement over listwise deletion; especially at larger percentages of missing data (20% and higher). At larger percentages of MNAR data, listwise deletion is a very poor choice and should therefore be used cautiously if a MNAR mechanism is a possibility. Because our technique is easy to implement for discrete variables, the authors encourage its use in cases where MNAR is suspected.

In addition, our original goal of developing a method to help to determine the missing-ness mechanism (MAR or MNAR) seems plausible from these results. Further testing will be needed to confirm the effectiveness of this method, but it does seem reasonable that if multiple data sets exist, each with varying levels of missing data, this technique could be used to plot the goodness of fit for each data set and compare to the goodness of fit using listwise deletion. A pattern may emerge similar to what we see in Figure 6 for either MAR or MNAR data. If only one data set is present, one could divide the data into various test sets and plot in a similar fashion. We will seek to test this method in future research.

7. LIMITATIONS AND FUTURE RESEARCH

This research has several limitations. First, the authors wish to test the technique with more test sets/BNs, as stated above, as well as against methods other than listwise deletion – multiple imputation in particular. Second, while our method works well for discrete values –*'Yes', 'No', 'High', 'Low', 'Medium'*, determining how to retain the missing state would be more difficult when data are continuous – temperature for example. '0' is a valid temperature and therefore could not be used as a marker for missing-ness, therefore the *'nr'* marker may require modifications to the BN learning algorithms. The authors do believe that this is a promising method for incorporating missing data that is supposed or known MNAR. It retains the underlying correlations among the data and is easy to implement for discrete sets.

REFERENCES

Meltem Almedar. 2009. A Monte Carlo Study: The Impact of Missing Data in Cross-Classification Random Effects Models. *Educational Policy Studies Dissertations*. Paper 34.

Ingo Beinlich, HJ Suermondt, RM Chavez, GF Cooper. 1989. The ALARM Monitoring System: A Case Study with Two Probabilistic Inference Techniques for Belief Networks. *Proceedings of the 2nd European Conference on Artificial Intelligence in Medicine*, pp. 247-256. Springer-Verlag.

Robert Cowell, G. Dawid, S. Lauritzen and D. Spiegalhalter. 1999. Probabilistic Networks and Expert Systems. Springer-Verlag, New York, NY.

Arthur Dempster, N. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39(1):1–38.

Nicholas Horton and K.P. Klienman. 2007. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *American Statistics. 61*, 79-90.

Finn Jensen. 2001. Bayesian Networks and Decision Graphs. Springer-Verlag, New York, NY.

Erica Klarreich. 2015. In Search of Bayesian Inference. *Communications of the ACM*, Vol. 58 No. 1, 21-24.

S. Lauritzen and D.J. Spielgelhalter. 1988. Local Computation with Probabilities in Graphical Structures and Their Applications to Expert Systems. *Journal of the Royal Statistical Society* B, Vol.50, No. 2.

Yang Lee, L. Pipino, J. Funk, and R. Wang. 2006. Journey to Data Quality. The MIT Press. Cambridge.

J Lin and P. Haug. 2008. Exploiting Missing Clinical Data in Bayesian network modeling for predicting medical problems. *Journal of Biomedical Informatics 41*. 1-14.

Roderick Little and D. Rubin. 1987 Statistical Analysis with Missing Data. John Wiley & Sons, New York.

Anders Madsen, F. Jensen, U. Kjærulff, and M. Lang. 2005. The Hugin Tool for Probabilistic Graphical Models. *International Journal of Artificial Intelligence Tools,* 14(3):507-543.

Patrick McKnight, K.M. McKnight, S. Sidani, and A. Figueredo. 2007. Missing Data: A Gentle Introduction. Guilford Oress, New York.

Richard Neapolitan. 2004. Learning Bayesian Networks. Pearson Education, Inc, Upper Saddle River, NJ.

Kristian Olesen, S. Lauritzen. and F. Jensen. 1992. aHUGIN: A System Creating Adaptive Causal Probabilistic Networks. Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence, 223-229.

Marco Ramoni, and P. Sebastiani. 1999. Learning Conditional Probabilities from Incomplete Data: An Experimental Comparision. Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics, 260-265.

Marco Scutari. 2010. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software.*Vol 35, Issue 3. Software available from http://www.bnlearn.com.

Valerie and M. Valtorta. 2009. Towards a Method for Data Accuracy Assessment Utilizing a Bayesian Network Learning Algorithm. *ACM Journal of Data and Information Quality*. Vol 1 No. 3. Article 13 .

Monica Tremblay, K. Dutta, and D. Vandermeer. 2010. Using Data Mining Techniques to Discover Bias Patterns in Missing Data. *ACM Journal of Data and Information Quality*, Vol 2, No. 1, Article 2.