University of KwaZulu-Natal

# THE MANAGEMENT OF MISSING CATEGORICAL DATA: COMPARISON OF MULTIPLE IMPUTATION AND SUBSET CORRESPONDENCE ANALYSIS

**GILLIAN MARGARET HENDRY**

**2015**

# THE MANAGEMENT OF MISSING CATEGORICAL DATA: COMPARISON OF MULTIPLE IMPUTATION AND SUBSET CORRESPONDENCE ANALYSIS

By

**GILLIAN MARGARET HENDRY**

Submitted in fulfilment of the academic

requirements for the degree of

Doctor of Philosophy

In

Applied Statistics

in the

School of Mathematics, Statistics and Computer Science

University of KwaZulu-Natal

Westville Campus

November 2015

# DEDICATION

To my husband, Keith, and my children, Neil and Liesl,

for allowing me the indulgence of these past few years

to enjoy the stimulation and realise a dream

# ABSTRACT

Missing data is a common problem in research and the manner in which this 'missingness' is managed, is crucial to the validity of analysis outcomes.

This study illustrates the use of two diverse methods to handle, in particular, missing categorical data. These methods are applied to a set of data which intended to identify relationships between asthma severity in children and environmental, behavioural, genetic and socio-economic factors. This dataset suffered from substantial missingness.

The first method involved the application of two approaches to multiple imputation, each adopting different distributional specifications. A practical challenge, previously undocumented, was encountered in the application of multiple imputation when interactions, to be identified and included in the analysis model, were needed for the imputation model. This study found that by imputing a single set of complete data using the expectation maximization (EM) algorithm for covariance matrices, it was possible to identify relevant interactions for inclusion in the imputation model.

The second method illustrated the application of correspondence analysis to a subset of the data that includes only the measured data categories. The application of subset correspondence analysis (s-CA) with incomplete data, as well as its sensitivity to the type of missingness, has not been well documented, if at all. There is also no evidence of research in which interactions have been added to an analysis with s-CA. In this study its use, both with and without interactions, was illustrated and the results, when compared to those from the multiple imputation approach, were found to be similar and favourably complementary. A simulation study found that s-CA performed well with any type of missingness, provided the amount of missingness is less than 30% on any variable with incomplete data.

Across all analyses, relationships found between asthma severity and factors were consistent with known relationships, thus providing confirmation of the reliability of the methods.

# DECLARATION

The research work described in this thesis was carried out in the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville campus, from January 2010 to December 2015, under the supervision of Professors Temesgen Zewotir, Rajen Naidoo and Delia North.

I, Gill Hendry, declare that this thesis is my own, unaided work. It has not been submitted in any form for any degree or diploma to any other tertiary institution. Where use has been made of the work of others, it is duly acknowledged in the text.

November, 2015

_____ _____

Mrs. Gillian Hendry Date

_____ _____

Professor Temesgen Zewotir Date

_____ _____

Professor Delia North Date

_____ _____

Professor Rajen Naidoo Date

# PUBLICATIONS

The following papers have been published from this thesis.

1 Hendry G, North D, Zewotir T, Naidoo RN. The application of subset correspondence analysis to address the problem of missing data in a study on asthma severity in childhood. *Statistics in Medicine* 2014, 33(22): 3882 – 3893

2 Hendry GM, Naidoo RN, Zewotir T, North D, Mentz G. Model development including interactions with multiple imputed data. *BMC Medical Research Methodology* 2014, 14:136

The following papers have been submitted for publication to journals.

1 Hendry GM, Zewotir T, Naidoo RN, North D. The effect of the mechanism and amount of missingness on subset correspondence analysis. *Correspondence in Statistics*. Under review

2 Hendry GM, Zewotir T, Naidoo RN, North D. A review on the application of multiple imputation and subset correspondence analysis in the South Durban health study. *BMC Medical Research Methodology.* Under review

# CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ACKNOWLEDGEMENTS

I am grateful to my supervisors, Professor Temesgin Zewotir and Professor Delia North, for their patience, guidance, advice and encouragement throughout this process. Also, to Professor Rajen Naidoo, I wish to express my deep appreciation for supplying the data for this study, and for his supervision and advice as well as his invaluable input into revisions of all the manuscripts. I have learnt a great deal from you all.

I would like to thank friend and mentor, Professor Paul Fatti, for his willingness to listen and for his invaluable advice which resulted in the ultimate direction of this study.

To Professor Michael Greenacre and Professor John Graham, authorities in your respective fields, thank you for your generous sharing of knowledge and advice.

Grateful thanks go to my husband, Keith, and children, Neil and Liesl, for their patience and support throughout the many years of this study. Without your understanding this would not have been possible. To Neil, special thanks for your advice and help with so many aspects on the computer and technical side. Special thanks also to Liesl, for sharing your writing expertise at all stages of this study and for being my listening board.

To my sister, Ethel Ross, a big thank you for doing an amazing job on the proof reading.

To my Mum, who was always so interested to hear about what I was studying and followed the progress each and every day, thank you for caring.

Finally, to my extended family and friends, a sincere thank you for your patience and encouragement over the past years and for showing an interest in what I was doing.

# ABBREVIATIONS

| | | |
|---|---|---|
| A1 | - | age 1: 8 - 9 years |
| A2 | - | age 2: 10 years |
| A3 | - | age 3: 11 years |
| A4 | - | age 4: 12+ years |
| ANOVA | - | analysis of variance |
| ASMI | - | asthma severity: mild intermittent |
| ASMP | - | asthma severity: mild persistent |
| ASMS | - | asthma severity: moderate to severe |
| ASN | - | asthma severity: no asthma |
| ASNI | - | asthma severity: none or mild intermittent |
| Bd | - | breakfast: daily |
| BMI | - | body mass index |
| Bnd | - | breakfast: not daily |
| BW1 | - | birth weight 1: up to 2.5 kg |
| BW2 | - | birth weight 2: > 2.5 kg |
| C | - | perceived weight: correct |
| CA | - | correspondence analysis |
| CC | - | complete case |
| COR | - | relative contribution to inertia |
| CPS | - | current population survey |
| CTR | - | absolute contribution to inertia |
| DA | - | data augmentation |
| df | - | degrees of freedom |
| DN | - | north Durban |
| DS | - | south Durban |
| e | - | stove: electric |
| E1 | - | exercise 1: < twice a week |
| E2 | - | exercise 1: 2 – 4 times a week |
| E3 | - | exercise 1: > 4 times a week |
| EM | - | expectation maximization |
| ETS | - | environmental tobacco smoke |
| FCS | - | fully conditional specification |

| | | |
|---|---|---|
| Fe | - | food availability: enough |
| FEM | - | gender: female |
| FIML | - | full information maximum likelihood |
| FNBd | - | Interaction: fear –no * breakfast – daily |
| FNBn | - | Interaction: fear –no * breakfast – not daily |
| Fne | - | food availability: not always enough |
| FrN | - | fear: no |
| FrY | - | fear: yes |
| fSVN | - | Interaction: gender – female * smoke in vehicle – no |
| fSVY | - | Interaction: gender – female * smoke in vehicle –yes |
| FYBd | - | Interaction: fear –yes * breakfast – daily |
| FYBn | - | Interaction: fear –yes * breakfast – not daily |
| g | - | stove: gas |
| I1 | - | income 1: up to R1000 |
| I2 | - | income 2: R1001 – R4500 |
| I3 | - | income 3: R4501 – R10000 |
| I4 | - | income 4: R10001 + |
| JAV | - | just another variable |
| M% | - | percentage missing |
| MAL | - | gender: male |
| MAR | - | missing at random |
| MCAR | - | missing completely at random |
| MI | - | multiple imputation |
| ML | - | maximum likelihood |
| MM | - | missingness mechanism |
| MNAR | - | missing not at random |
| mSVN | - | Interaction: gender – male * smoke in vehicle – no |
| mSVY | - | Interaction: gender – male * smoke in vehicle –yes |
| MVNI | - | multivariate normal imputation |
| n | - | stove: none |
| N1 | - | number of people 1: 1 – 4 people |
| N2 | - | number of people 2: 5 – 7people |
| N3 | - | number of people 3: > 7 people |
| NNN | - | Neo-natal care: no |

| | | |
|---|---|---|
| NNY | - | Neo-natal care: yes |
| O | - | perceived weight: overweight |
| OR | - | odds ratio |
| p | - | stove: paraffin |
| PN | - | pet: no |
| PY | - | pet: yes |
| QLT | - | quality |
| S1 | - | strategy 1 |
| S2 | - | strategy 2 |
| s-CA | - | subset correspondence analysis |
| SEM | - | structural equation modelling |
| SES | - | socio-economic status |
| SN | - | smokers in the home: no |
| SPN | - | smoked while pregnant: no |
| SPSS | - | statistical package for the social sciences |
| SPY | - | smoked while pregnant: yes |
| SVN | - | smoke exposure in vehicles: no |
| SVY | - | smoke exposure in vehicles: yes |
| SY | - | smokers in the home: yes |
| T1 | - | TV 1: < I hour a day |
| T2 | - | TV 2: 1 - 3 hours a day |
| T3 | - | TV 3: > 3 hours a day |
| TI%FULL | - | total inertia as a percentage of the full inertia |
| TOTINR | - | total inertia |
| U | - | perceived weight: underweight |
| VLBW | - | very low birth weight |
| VN | - | violence: no |
| VY | - | violence: yes |
| WN | - | weapons: no |
| WWN | - | work and wear: no |
| WWY | - | work and wear: yes |
| WY | - | weapons: yes |

# Chapter 1

# INTRODUCTION

The problem of missing data is unavoidable and has hindered researchers from the time field research began. It is especially common in the medical and social sciences (Rubin, 1996). Analytic procedures that were developed early in the twentieth century were designed to be used with complete data sets (Graham, 2009) and thus the presence of missing data presents a challenge.

When data are missing, researchers need to proceed cautiously and be aware that the reason for their missingness as well as the manner in which it is addressed may produce bias which can impact the results and ensuing inferences. This can potentially affect the validity of research outcomes – a fact that is often ignored in medical literature (Wood et al., 2004).

Early researchers adopted a number of methods to deal with the missing data. These methods included, amongst others, mean substitution, case-wise deletion, hot deck imputation and the indicator method. These were in many cases merely a means to bypass the problem of missing data and were seldom successful in achieving the primary goal of analysis – to obtain unbiased estimates of population parameters. There is generally widespread agreement regarding the strengths and weaknesses of these methods.

Howell (2007) noted that modern techniques such as maximum likelihood and multiple imputation procedures have come far in narrowing the gap between the ideal and the practical. He further suggested that, in specialized areas like the treatment of missing data, it takes time for newer methods to be understood and adopted, as confirmed by the low usage of these methods, as evidenced in the literature.

## 1.1 Missing data classifications

Missing data are traditionally categorised into three different types of missingness: missing completely at random (MCAR); missing at random (MAR) and missing not at random (MNAR). The type of missingness is determined on the basis of the process that led to the missingness. This terminology for missingness was introduced by Donald Rubin (1976) , one of the pioneers of missing data research.

### 1.1.1 Missing completely at random

Data are 'missing completely at random' (MCAR) if the probability of an item missing has nothing to do with the value of the item or with the value of any other variable in the data. In other words it is a completely random happening with missing values being randomly distributed across all observations. Thus one observation of a variable is as likely to be missing as any other. MCAR data can result from, for example: someone missing an interview session due to traffic problems; accidentally turning two pages of a questionnaire instead of one; or a respondent genuinely not knowing the answer to a question. This missingness is easy to handle and does not cause bias in the results. However, depending on how the missing data is managed, MCAR can result in a loss of statistical power.

### 1.1.2 Missing at random

Data are 'missing at random' (MAR) if the probability that a variable is missing depends on another variable that may or may not be part of the data set. For example, the probability that income is recorded may depend on the gender of the respondent. It could be that females are more willing to divulge income than males. Provided the variable with which the missingness is related – in this case gender - is fully recorded, the missing data can be considered MAR. If the variable with which the missingness is related contains some non-response itself, then the data are not MAR.

Little and Rubin (1987) refer jointly to data that are MAR or MCAR as 'ignorable'. When the cause of missingness is known, measured and available for use in analysis, it is also called an *accessible missing data mechanism* (Graham and Donaldson, 1993). Correct application of this

cause of missingness in analysis results in the adjustment of biases associated with missing data. On the other hand, when the cause of missingness is either not measured or is related to the variable containing the missingness, it is referred to as an *inaccessible missing data mechanism*. This is equivalent to a 'non-ignorable mechanism' and is commonly termed 'missing not at random'.

### 1.1.3 Missing not at random

All data that are neither MAR nor MCAR are described as 'missing not at random' (MNAR). This is a difficult situation to deal with as, even though the cause of the missingness may be known to the researcher, it is not measured or available for use. In addition, missingness that is MNAR will yield biased parameter estimates. A familiar illustration, taken from the medical field, of MNAR data is where patients drop out of a study because of discomfort they experience as a result of a treatment they are receiving. If the discomfort factor is not measured for all patients then this missingness is 'not at random'. Another form of MNAR data occurs when the missingness depends on the missing value itself – for example, where people who consume excessive amounts of alcohol are reluctant to reveal their drinking habits. In the extreme case, where all refuse to divulge their excessive drinking habits, it is called *censoring*.

A recognised treatment of censoring and MNAR data that will result in unbiased estimates of the parameters of the model is to explicitly model the missingness. Unfortunately, this is a difficult exercise as it is often not clear what lies behind the missingness and furthermore the missingness depends on unobserved data. As the term *non-ignorable mechanism* implies, the analysis that is being carried out cannot be completed unless a model governing missingness is also able to be written (Howell, 2007).

## 1.2 Early contributions

One of the earliest apparent applications of data imputation dates back to the 1940's at the United States Census Bureau. According to Scheuren (2005) this endeavour to impute missing data, which was at that time referred to as 'allocating' data, was applied to the 1940 Decennial Census. Hot deck imputation was used to fill in the missing items of data (Ford, 1983).

Realizing that this form of imputation introduced a degree of bias with an underestimation of variance, Donald Rubin – a pioneer of multiple imputation – was consulted to help with a solution to the problem. As a result, multiple imputation was conceived and first documented by Rubin in 1977 (Rubin, 2004a). Additional influential research relating to the subject was reported by several authors, including Rubin (1976), Dempster et al. (1977) and Heckman (1979). It was not until 1987 that dramatic changes were experienced in the field. Two important books (Little and Rubin, 1987, Rubin, 2004b) were published and these laid the groundwork for significant advances in the treatment of missing data. Coupled with these publications was the advent of powerful personal computers which led to the development of much needed software. Other important publications from 1987 described methods for dealing with missing data using structural equation (SEM) software (Allison, 1987, Muthén et al., 1987); and data augmentation that would later become a cornerstone of some multiple imputation software (Tanner and Wong, 1987).

Current software, mostly developed since 2000, includes both stand-alone software and routines incorporated into popular statistical packages. Two significant contributions were made in this regard by Schafer (1997) who developed **NORM** (Novo and Schafer, 2010) – a stand-alone Windows package – and Van Buuren and Groothuis-Oudshoorn (2011) who developed **MICE** – a routine implemented in both S-PLUS and R.

## 1.3 Managing missing data

Over the years a variety of methods have been used to 'handle' missing data but most of them do not, in essence, effectively deal with the missing data. Some are acceptable under certain conditions and others should not be used at all. Their success lies in the assumption of the type of missingness present as well as in the application of the method. Accessibility to techniques for managing missing data was, to begin with, restricted to what are commonly known as *ad hoc* methods. These include variable or case deletion and a number of single imputation methods. More recently, sophisticated techniques including multiple imputation have been developed to manage missing data. With increasing accessibility and the advancement of modern computers with superior computational capabilities, the use of these methods should become more evident in the literature.

### 1.3.1 *Ad hoc* methods

***Complete case analysis***

Probably the most commonly used and simplest approach to dealing with missing data is complete case analysis, also known as listwise deletion, casewise deletion or available case analysis. Listwise deletion excludes all cases with any missing item(s) from the analysis. Consequently there can be a substantial decrease in the number of cases included in the analysis which results in a loss of power. While applying this method to data that are MCAR generally leads to unbiased parameter estimates, this may not be the case if the cases with missing values differ significantly from the complete cases., It has been suggested that if the data are not MCAR, bias will result in the parameter estimates (Howell, 2007). Contrary to this, Graham (2012) reported that 'when missingness is MAR, regression coefficients for pre-test variables predicting post-test variables will often be tolerably unbiased'.

Despite the above reservations, Howell (2007) suggests that this method may, even in the case of data missing not at random, be better than many of the alternatives. Graham (2009) believes that this method can still be useful, especially if the loss of cases due to missing data is less than about 5%, since both biases and loss of power are likely to be negligible.

***Pairwise deletion***

In pairwise deletion, cases are excluded from any calculation involving variables for which they have missing data. Although this method does make use of the largest possible sample to estimate parameters, different calculations are based on different sets of data, with different sample sizes and different standard errors. Furthermore, unless missingness is MCAR, bias in the parameter estimates may still exist. It is also possible, and in fact not uncommon, that the correlation or covariance matrices resulting from this deletion will not be positive definite thus preventing analysis from being completed (Howell, 2007). A further limitation is that standard errors cannot be estimated because the sample size is needed for this calculation and this value is not constant with pairwise deletion. This method is not recommended as a general solution to missing data (Graham, 2009, Howell, 2007). A variation of pairwise deletion is what is called complete variable analysis. In this case, variables are omitted from analysis if they suffer from missingness. This can result in a potentially important predictor being omitted from analysis.

Both of the methods described above lose potential valuable information because of the special selection of cases or variables.

Instead of decreasing the size of the data set by removing incomplete cases or variables, many methods have been developed whereby missing data are imputed. These methods range from extremely simple to rather complex. Some methods do single imputations, i.e. each variable with missing observations is imputed independently of the next. Other methods impute missing data across several variables simultaneously.

***Hot deck imputation***

Hot deck imputation was first used in 1947 at the U.S. Census Bureau for item non-response in the Income Supplement of the Current Population Survey (CPS) (Andridge and Little, 2010).

For a case with single or multiple missing values, a similar complete case, based on the observed responses, is selected. The missing values are then replaced by the corresponding values from the complete case. This works reasonably well providing the missing data are kept to a minimum. As non-response increases and more data are replaced, the parameter estimates and their standard errors are compromised. Although hot deck imputation may be useful in some circumstances, it is not commonly used today. It has been suggested that while hot deck has some 'attractive features', it underestimates the variance and its 'bias reduction properties are suspect, at best, without some form of supplementation' (Scheuren, 2005).

***Mean substitution***

As the name suggests, mean substitution involves replacing the missing value by the mean of the observed values for that variable. This method of imputation results in an underestimation of the standard error as, in effect, no new information has been added to the data. It is generally thought that this method should not be considered (Graham, 2009, Howell, 2007). It has even been suggested that this method is 'the worst of all possible strategies' for dealing with missing data (Graham, 2012).

***Missing value coding***

One idea that was made popular in the behavioural sciences by Cohen et al. (2003) was the addition of indicator variables to code for missing items while replacing missing data with mean values. This method, however, results in biased parameter estimates (Jones, 1996) and is no longer recommended (Allison, 2002, Graham, 2009, Howell, 2007). A similar method is applied to categorical data when a single categorical covariate has missing data. In this case, a separate category is introduced for the missing data. This, too, has been shown to lead to significant bias (Vach and Blettner, 1991).

***Last Observation Carried Forward (LOCF)***

Specific to longitudinal data problems, where multiple measurements are taken per subject over time, this method replaces a missing measurement with the last measurement taken. Whether for a multi-stage longitudinal analysis, where a subject drops out of the study before completion, or with single time point analysis, biased results and an underestimation of variability will result. This will negatively impact on precision and hence inferences may not be valid. This is true even under MCAR. Molnar et al. (2008) report that this method 'provides no benefits', and 'creates unnecessary risk of generating biased or even false conclusions'. Because the researcher is making assumptions about a subject had they not dropped out, this method is not recommended (Howell, 2007).

***Regression-based single imputation***

A slight improvement on mean substitution, regression-based substitution uses additional information about the case to impute the missing value. Howell (2007) suggests that this works reasonably well providing the variable with the missing data is strongly related to the other independent variables. The higher the correlation between the missing variable and the predictor variables, the better will be the imputation (Graham, 2012). It has even been described as the 'best of the simple solutions to missing data' (Lynch, 2003). However, because the missing value is replaced with a value predicted from the other variables, new information is not added but rather the sample size is increased and the standard error is underestimated. This will, in turn, affect the resulting regression coefficients. According to Graham (2012), even though the concept is a sound one, he does not recommend its use in general. It is important

to note, nonetheless, that regression-based imputation forms the basis for many of the modern and highly recommended missing data methods.

All of the methods described above in which there is *ad hoc* deletion or replacement of missing data are appealing in that they are easy to implement. In addition, some of them are readily available in most statistical software packages. Graham (2012) suggests, however, that none of these methods deals effectively with missing data but they instead circumvent the problem so that some further analyses can be attempted. There is general consensus that care should be taken when using these *ad hoc* methods as they have been shown to have serious drawbacks (Little and Schenker, 1995, Schafer and Graham, 2002). Graham et al. (2003) refer to them as "unacceptable methods" and in a later publication, Graham (2009) encourages researchers to use modern missing data procedures that are known to be good and that will produce unbiased results even when the data are MAR.

### 1.3.2 Modern methods

Two more conventional methods, both based on strong statistical principles and recommended to deal with missing data, are maximum-likelihood (ML) and multiple imputation (MI) procedures. Graham (2009) believes that, compared to over 25 years ago, these methods are 90% of the way to reaching the 'hypothetical ideal' in terms of missing data solutions. Simulation studies have shown that ML algorithms produce superior results to the traditional *ad hoc* methods when dealing with missing data (Arbuckle et al., 1996, Enders and Bandalos, 2001, Muthén et al., 1987). There is also much support for the use of MI (Azur et al., 2011, Desai et al., 2011, Donders et al., 2006) and Scheffer (2002) even suggests that 'multiple imputation is always better than case deletion or single *ad hoc* methods'.

#### *Maximum likelihood*

A number of ML algorithms for use with missing data are available. In each case, multivariate normality is assumed. Possibly the most commonly used of these is the full information maximum likelihood (FIML) estimation method, which is implemented in the AMOS software (Arbuckle, 2006). In this process, a single case of raw data is read in and, using the available information in the case, the ML function is maximized. An overall estimate of the ML function is then obtained by summing across the individual cases. In this procedure the missing data

and the parameter estimation are dealt with in a single step. FIML is thought to provide excellent parameter estimates as well as reasonable standard errors (Enders, 2001, Graham, 2012).

***Multiple imputation***

Adding random error to overcome the problem of underestimation of standard errors is a key factor in all variations of MI. Several algorithms have been developed to perform MI. The two most widely used approaches are multivariate normal imputation (MVNI) and fully conditional specification (FCS). Both are iterative techniques and produce multiple (m) complete data sets. Estimates for the missing data are obtained by regressing the incomplete variable on all other variables in the model. While MVNI assumes all variables follow a multivariate normal distribution, FCS is less restrictive and tailors the regression model to suit the type of variable to be imputed.

The MVNI algorithm is adopted by the NORM software (Schafer, 1999) while FCS is implemented in the MICE routine in S-Plus and R (Van Buuren and Groothuis-Oudshoorn, 2011).

Each of the m data sets produced at the imputation stage using both MVNI and FCS is analysed using the analysis of choice. The resulting parameter estimates are then combined using Rubin's rules (Rubin, 2004b) to yield the point estimate of the parameters and the MI-based standard errors. It has been suggested that, 'until something even better comes along', it is 'likely that MI will be the solution of choice' for the treatment of missing data (Howell, 2007).

Under the assumption of MAR, these MI and ML methods are specifically designed to provide unbiased parameter estimates and, even if this assumption is not met, these methods will always perform at least as well as the older *ad hoc* methods (Graham, 2012). It has also been suggested that, while specific non-ignorable methods to manage MNAR data (Demirtas and Schafer, 2003, Demirtas, 2005, Little, 1995) may be very useful, it is not necessarily true that these methods will be better than MI or ML methods for any particular empirical study (Graham, 2009). If the missingness model and its assumptions are incorrect, the MNAR model may perform worse than standard MAR methods (Demirtas and Schafer, 2003). In a study

investigating the effect of different imputation methods and different missingness mechanisms on the mean and standard deviation it was found that MI produced good results for MNAR data at missingness levels of less than 25% (Scheffer, 2002). In another study comparing different imputation techniques to deal with missing data in a multi-question depression scale, Shrive et al. (2006) found that MI performed well even with MNAR data. They report that this agrees with findings from Faris et al. (2002) who demonstrated a similarly strong performance by MI on MNAR data. It has also been shown that with both MI and ML, by including in the imputation model auxiliary variables – variables that are correlated with the missingness – the impact of non-ignorable missingness is reduced (Collins et al., 2001, Graham, 2009, Schafer, 1997).

Not only is it important to include auxiliary variables in the imputation model, but all variables to be used in the analysis model, including the dependent variable and interactions, should be included. The exclusion of any variable to be used at the analysis stage will result in biased estimates (Graham, 2009).

## 1.4 How researches are managing missing data

Even though significant advancements in the management of missing data and availability of appropriate software have been evident since 2000, the adoption of these recommended methods is slow. In a review of literature on PubMed from January 2000 to December 2009, which included all cohort studies with a sample size of at least 1000, the authors highlighted the 'continuing use of inappropriate methods to handle missing data' with only 7% of the studies using a recommended method for dealing with the missing data (Karahalios et al., 2012).

Another article on the use of MI in epidemiologic literature (Klebanoff and Cole, 2008) searched four leading epidemiologic journals for articles published from January 2005 to December 2006. Of the 99 articles containing the text '*imput*', only 12 used MI, while a further four used other acceptable methods. The authors of this review article expressed surprise at 'how infrequently multiple imputation appeared in epidemiologic manuscripts given the well-described shortcomings of simpler approaches'. In another review of 262 studies from three

leading epidemiologic journals, all published in 2010, it was found that only 13% of the studies used recommended methods such as MI and ML estimation. Complete case analysis was performed in 81% of the studies, despite the fact that the average proportion of missing data exceeded 10% (Eekhout et al., 2012).

A search on PubMed for articles relevant to this study was carried out. All studies published from 1 January 2012 to 31 July 2013 on associations between childhood asthma and various environmental factors, were included. Of the 50 studies that satisfied the inclusion criteria, 27 of them were cross-sectional. 89% of these suffered from missing data. The search revealed that not one of these cross-sectional studies indicated the use of MI or any other acceptable method of dealing with the missing data. Of the remaining 23 non-cross-sectional studies, only one indicated the use of MI to deal with the missing data.

## 1.5 The objectives of this study

This study addresses several aspects of analysis to deal with missing categorical data. Some MI algorithms are explored with a special investigation into the inclusion of interactions when data are incomplete. Another entirely different approach is also introduced that not only manages the missing data but is also unrestricted by the distributional requirements of the aforementioned MI methods. An investigation into the effect of the different missingness mechanisms on this method is also presented. These applications are carried out on a set of data from a study on childhood asthma in which the severity of asthma as well as environmental, behavioural, genetic and socio-economic factors are measured.

The main objectives of the study, which are addressed in specific papers, follow.

### 1.5.1 Research objective 1

The inclusion of interactions in the imputation model is not always straight forward. There are two schools of thought on how this can be done. One way is to impute the individual variables first and thereafter form the interaction product terms using the imputed data. This is called ‘passive’ imputation. Alternatively, the interaction product terms can be included in the

imputation model as additional variables – termed 'JAV' (just another variable) (White et al., 2011) – and imputed along with the other variables. It has been shown that with passive imputation, while the interaction terms are compatible with the variables, parameter estimates in the analysis model are affected by bias (Von Hippel, 2009). On the other hand, treating the interaction terms as JAV will produce interactions that are not always compatible with the individual variables; but the resulting parameter estimates and standard errors are unbiased.

Clearly, including the interactions as JAV at the imputation stage should be the method of choice if bias is to be avoided. This can present a challenge in practical terms since interactions are not always known *a priori* and need to be identified from the data. If the number of variables in the data is small, then all possible interactions can be included at the imputation stage. However, as the number of variables increases, this becomes impractical and in many cases computationally impossible.

While studies have been done on MI with interactions included (Desai et al., 2011, White et al., 2011), in all these cases the interactions are known prior to imputation. The dilemma that the researcher faces is: complete data is needed to identify interactions; but the interactions are needed to carry out the imputations. Addressing this practical challenge is thus one of the objectives of this study.

### 1.5.2 Research objective 2

Managing missing data with MI involves fitting the data to a model. This application is often restricted by complexities of models and distributional requirements. Furthermore, many of the MI algorithms are more suited to dealing with missingness in continuous variables, even though, in some applications, adaptations can be made to impute categorical variables and are found to work well (Graham, 2012).

The second objective of this study is to investigate the application of subset correspondence analysis (s-CA) to address the issue of missingness in the analysis of categorical data. This method of analysis adopts a very different approach to the more traditional aforementioned

method in that it fits a model to the data and there are no distributional restrictions to consider.

### 1.5.3 Research objective 3

While many studies have examined the effect of missingness mechanisms and the amount of missingness on MI (Hardt et al., 2012, Marshall et al., 2010, Peyre et al., 2011, Shrive et al., 2006) , it is not known what effect these factors might have on s-CA. The third objective is then to investigate the impact of the amount and mechanism of missingness on the application of s-CA as a solution to missing data management problems.

### 1.5.4 Research objective 4

This study illustrates the use of both MI and s-CA to manage the missing data when analysing relationships between variables. What is not clear is whether these two diverse methods arrive at the same conclusions with regard to relationships between variables that suffer from missingness. Moreover, to the knowledge of the researcher, the inclusion of interactions with s-CA has not been demonstrated in the literature. The fourth objective is to compare outcomes from these two methods while at the same time demonstrating how interactions can be added to an s-CA analysis.

### 1.5.5 Research objective 5

Many studies have been done on the effect of the environment, socio-economic status, genetics and behavioural patterns on asthma in children. Relationships between asthma and these factors are well documented and presented in Chapter 2. While the vast majority of these studies are internationally based, very little is known about the impact of these factors on the respiratory health of children in South Africa. The fifth objective is therefore the intrinsic aim of this study which is to identify the factors that affect the severity of asthma in children specific to the Durban South basin.

# Chapter 2

# The Data

## 2.1 Background literature review on asthma in children

Respiratory diseases are the most common cause of illness in children. Outdoor and indoor air quality, poverty, poor housing, malnutrition and poor medical services are contributory causes to the burden of illness from asthma and other respiratory problems in children (Ernst et al., 1995, Gold and Wright, 2005, Litonjua et al., 1999, Peden, 2003). Epidemiological and clinical studies link respiratory problems with unfavourable housing and living conditions (Rauh et al., 2008) (Rosenstreich et al., 1997, Williamson et al., 1997).

Laboratory and population-based studies have shown associations between stress experiences and asthma expression (Subramanian and Kennedy, 2009, Wright, 2008). In a study across cities in Los Angeles County, it was found that childhood asthma and community violence, as measured by assault hospitalisations, were significantly associated (Jeffrey et al., 2006) . A Canadian study on the association between asthma prevalence in school children and neighbourhood stressors, found that the stress in early childhood on children living in high crime neighbourhoods was associated with either the development of asthma or the worsening of symptoms (Pittman et al., 2012). Another study on the effect of community violence on childhood asthma was conducted in Brazil (Alves et al., 2012). This study investigated the degree of exposure to community violence and its effect on the respiratory health of the children. It was found that children that were more exposed to violence, including gang warfare and drug trafficking, showed higher asthma prevalence compared to non-exposed children. They further showed that children exposed to the maximum level of violence were nearly twice as likely to present asthma symptoms. For those who knew someone that had been either beaten or injured with a firearm or knife, there was a nearly 40% higher prevalence of asthma. In a study conducted in Vancouver, Canada, violence in the neighbourhood and lower socio-economic status (SES) of neighbourhoods were associated with asthma morbidity (Chen et al., 2007).

Even though SES has been linked to childhood asthma in different regions around the world, the findings cannot be reproduced in all regions (Von Mutius, 2000). In Britain, Singapore and Hong Kong, it was found that parent-reported asthma is more prevalent among subjects of a higher socio-economic status (Goh et al., 1996, Kaplan and Mascie-Taylor, 1985, Lai et al., 1996, Peckham and Butler, 1978); while studies in the USA have found that asthma prevalence is associated with poverty (Litonjua et al., 1999, Persky et al., 1998). Poyser et al. (2002) , in their study on socio-economic deprivation and asthma prevalence and severity in young adolescence in Cape Town, South Africa, found that pupils living in 'better-off areas' reported a higher prevalence of ever having had asthma than those living in the poorer areas. In contrast, they found that greater severity of the disease was associated with lower socio-economic groups. Von Mutius (2000) suggests that the significant differences in asthma prevalence between regions of similar ethnic backgrounds are very likely strongly influenced by the environment.

Later studies confirm this contradiction of the link between SES and development of asthma (Gold and Wright, 2005, Hancox et al., 2004, Kozyrskyj et al., 2010, Shankardass et al., 2007). It has been suggested that 'SES is a rough marker of a variety of environmental/behavioural exposures' including, amongst others, dietary habits, family size, access to health care, environmental tobacco smoke (ETS) and allergen exposure (Forno and Celedón, 2009). Kozyrskyj et al. (2010) suggest that the contradictory findings regarding the association between SES and childhood asthma may be a function of the variable(s) used to measure SES.

Studies on the association between family size and asthma prevalence show conflicting results. While many studies have found a negative relationship between family size and asthma prevalence (Dik et al., 2004, Karmaus and Botezan, 2002, McKeever et al., 2001), others have found that asthma prevalence is higher in larger families (Davis and Bulpitt, 1981) and others still have found no relationship between these variables (Nafstad et al., 2005). In a study on medical records of more than half a million 17 year olds in Israel who had ever had asthma, it was found that in families with more than three children, asthma prevalence was inversely related to the number of children in the family (Goldberg et al., 2007).

A study on the association between household income and asthma symptoms among elementary school children in Seoul found that income and asthma were inversely associated

after adjusting for other potential risk factors (Choi et al., 2012). It further showed that this association was modified by the number of siblings. While there was no significant effect of income on asthma symptoms for children with two or more siblings, low income was still a significant factor for children with fewer than two siblings (OR 1.41; 95%CI, 1.09 – 1.81).

According to Bae et al. (2008), studies confirm that insufficient food intake can result in asthma. In a community childhood hunger project, conducted in Washington DC, it was found that 'poor hungry children were more likely than poor but not hungry children' to suffer from health problems such as asthma (Alaimo et al., 2001).

Domestic air quality has also been found to be linked to asthma. Airborne allergens, shown to exacerbate asthma, include dust mites, cockroaches and domestic cats and dogs (Gent et al., 2009). It has also been shown that asthma prevalence is greater in homes with wood and coal burning stoves compared to those using other sources of heating (Jones, 1998). More recently, Bates et al. (2013) conducted a study on the effect of cooking fuels on children in Nepal. They found that, compared to the use of electricity for cooking, there was increased incidence of lower respiratory infection with the use of solid fuels, including wood, coal and paraffin.

A review of 32 articles up until 1999 with regard to exposure to pets and the risk of asthma (Apelberg et al., 2001) yielded conflicting results. However, combining results from these articles it was found that the pooled risk for asthma in studies on a population with median age in excess of 6 years indicated a small but significant effect (OR 1.19; 95% CI 1.02 – 1.40). According to a more recent review of similar articles from 1966 to 2007 (Takkouche et al., 2008) , it was found that, on pooling results, there was evidence that while exposure to dogs slightly increases the risk of asthma (OR 1.14; 95% CI 1.01 – 1.29), exposure to cats has a slight preventative effect on asthma (OR 0.72; 95% CI 0.55 – 0.093). Exposure to furry pets of undetermined type was not conclusive. Carlsen et al. (2012), in their study of 11 European birth cohorts, investigated whether pet ownership in infancy leads to asthma or allergy at school age. They found that ownership of single types of furry pets or birds in the first 2 years of life neither increased nor decreased the risk of asthma in school-aged children. However, they did find that living with furry pets in the first 2 years appeared to reduce the likelihood of becoming sensitised to aeroallergens in early school age.

Goodwin and Cowles (2008) found an association between both pre- and post-natal exposure to cigarette smoking. This is consistent with the findings of DiFranza et al. (2004) who also reported that respiratory risk associated with parental smoking seems to be greatest during foetal development and the first few years of life. Rayens et al. (2008) found a decrease in asthma related ER visits since smoke-free laws were introduced. A study on environmental tobacco smoke (ETS) in cars (Sendzik et al., 2009), found that ETS is associated with a greater likelihood of asthma and other chronic lung diseases. Furthermore, risks for children exposed to ETS in cars are greater than those of children exposed to ETS in the home and they further showed that children are more susceptible to the effects of ETS exposure than is the case for adults. It would seem that because of the restricted area in a car within which smoke circulates, the levels of ETS in cars pose a significant risk to children. Results from a longitudinal study (Sly et al., 2007) found that by the age of 14, children exposed to ETS in cars are more likely to have a current or persistent wheeze, and decreased lung function compared to children who were not similarly exposed. They further found that the risk for children exposed to ETS in cars was greater than that of children exposed to smoking in the home.

A cross-sectional study on the effect of exposure to air stack emissions of sulphur dioxide from petroleum refineries on asthma among children aged 6 months to 12 years was carried out in Montreal, Canada (Deger et al., 2012). It was found that a significant relationship exists between exposure to refinery stack emissions of $SO_2$ and the prevalence of asthma (OR 1.14; 95% CI 0.94 – 1.39). These results concur with numerous other studies reporting an increased prevalence of asthma among children living in proximity to industrial areas, including refineries and petrochemical industries (Charpin et al., 1988, Henry et al., 1991, Maantay, 2007, Roberts and Ehrlich, 2009).

Approximately 250 occupational pollutants including chemicals, metals, enzymes, drugs, and others, are known as risk factors for asthma (Venables and Chan-Yeung, 1997). These pollutants and compounds usually occur in high concentrations in the work place. It has been suggested that prolonged exposure of the general population to lower concentrations of these compounds could initiate or exacerbate asthma in susceptible individuals (Becher et al., 1996).

The association between the prevalence of asthma in school children and passive smoking and obesity was carried out in Mexico (Bedolla-Barajas et al., 2013). They found that neither

obesity nor passive smoking, where one of the parents smoked, is significantly associated with asthma in children aged between 6 and 12 years.

There is an increasing awareness that very low birth weight (VLBW) children are at risk of experiencing long-term health problems (McManus et al., 2012). These children suffer, most commonly, from respiratory disease and are three times as likely to get asthma as normal birth weight children (Brooks et al., 2001, Hack et al., 2005).

In a study on the relationship between very low birth weight (VLBW) and the development of asthma (Mai et al., 2003), it was found that, at age 12, asthma was more frequently found among the VLBW children than the term children. They also reported a significant association between the VLBW children who received neo-natal care in the form of mechanical ventilation and/or oxygen supplementation, and those with a history of asthma by the age of 12.

According to Corbo et al. (2008), high body weight and spending a lot of time watching TV each independently increase the risk of asthma symptoms being present in children. In their investigations on the associations between asthma and wheeze in children and body mass index (BMI), sports, television viewing and diet, they found that subjects who spent 5 or more hours per day watching television, were more likely to experience wheeze or current asthma, compared with those who viewed TV less than 1 hour a day.

A study on the effects of elevated BMI among low birth weight children on asthma prevalence was conducted in Taiwan (Lu et al., 2012). Their results suggest that low birth weight predisposes one to develop asthma. They further showed that, amongst the low birth weight subjects, an elevated BMI in adolescence was associated with a higher risk for asthma. It was further found that low birth weight boys who were either normal weight or underweight adolescents, had an increased risk for developing asthma. The same association, however, was not true for girls.

While childhood asthma is generally associated with younger children (Asher et al., 2006), it is more prevalent in boys, than in girls, before puberty. In fact, some investigations have shown male: female ratios for asthma prevalence to be as high as 4:1 (Bonner, 1984). Hospitalisation

rates for asthma reflect the difference in asthma severity between boys and girls. One study on asthma-related hospital admissions by age and sex in Finland shows that at age 1 year, hospital admission rate is 5.3/1000 for boys and 2.9/1000 for girls. These rates begin to equalize at puberty and remain similar throughout adolescence (Harju et al., 1996). A study in the USA on children of 2 – 11 years of age found that, in girls, asthma prevalence increased approximately linearly from 12% among underweight girls to 33.3% among overweight girls. In boys, asthma prevalence was 36.4% for those underweight, 19.1% for normal weight boys and 34.8% among overweight boys (Kwon et al., 2006). Another study examined the influence of pre-natal exposure to particulate air pollution on the respiratory health of full-term children at 6 years of age (Hsu et al., 2015). It was found that increased exposure at 16-25 weeks gestation was associated with early development of asthma. Further it was shown that this association was true only for boys.

## 2.2 The survey and descriptive statistics

In 2004 the eThekwini Municipality in KwaZulu-Natal, South Africa commissioned researchers from the University of KwaZulu-Natal (UKZN) to conduct a study on the effect of ambient air pollution on the respiratory health of children in the South Durban region (Naidoo et al., 2013). A highly industrialized area, the South Durban Industrial Basin is recognised as one of the worst polluted areas in Southern Africa (Matooane and Diab, 2001). It is home to large crude oil refineries, petrochemical plants, a paper mill, a bulk chemical storage facility, motor manufacturing plants and many other 'smoke stack' industries.

Data were collected from seven schools in two regions – four schools in the industrialised Durban South region and a further three schools, from similar socio-economic backgrounds, in the non-industrialised Durban North region. In order to achieve a sample of persistent asthmatics with adequate power to determine association between asthma and variables of interest, all students from Grades 3 – 6 completed a 'screening' questionnaire to determine the status of their respiratory health with specific reference to asthma and asthma symptoms. The study sample comprised one or two randomly selected Grade 4 classes from each school as well as additional children from Grades 3 – 6 identified as having persistent asthma on the basis of the 'screening' questionnaire. Of the 422 children from the randomly selected classes,

342 agreed to participate in the study. A further 93 children across all grades were identified as having persistent asthma and were invited to join the study. Of these, 81 participated in the study. The total study sample thus consisted of 423 children.

Data covering socio-economic, genetic, behavioural and environmental aspects were gathered from the children, their guardians and their families by means of four surveys administered by trained interviewers from the research team. 'Caregiver', 'adult' and 'family' interviews were conducted with family members at home while the 'child' interview was carried out at school.

When preparing the data for this study, it was decided that any subject without an asthma classification would be excluded. In addition, it was found that, for a number of subjects, demographic information across the four surveys was not consistent and, because these records were deemed unreliable, they were also excluded from the final data set. In all 41 subjects were excluded, thus leaving a final sample of 382 children for inclusion in this study.

From the variables collected across the four surveys, 21 environmental, genetic, socio-economic and behavioural variables as well as the four-tiered asthma severity variable were chosen to be included in this study. Careful selection was made to ensure that each of the environmental, genetic, socio-economic and behavioural constructs was adequately represented by several variables that are well studied as factors affecting asthma. Details of these variables, as well as their frequencies, are presented in Table 2.1.

Most of the variables used are self explanatory. However, some need further description to clarify their use in this study. The survey questions that lead to these variables are outlined below.

The variables 'age' and 'gender' apply to the child in the study. Also concerned with the child are the variables 'exercise' (How many times per week do you play or exercise enough to make you sweat and breathe hard?); 'TV' (About how many hours did you watch TV yesterday?); 'fear' (Are you afraid you will be hurt by violence in your neighbourhood?); 'breakfast' (How often do you eat breakfast – every day, on some days, rarely, never, or on weekends only?); 'perceived weight' (Do you consider yourself to be overweight, underweight or about the right

weight?); ‘violence’ (While you have lived in your neighbourhood, has anyone ever used violence, such as in a fight (hitting, pushing, shoving), against you or any member of your family anywhere in your neighbourhood or home?); ‘weapons’ (Was there a fight in which a weapon was used anywhere in your neighbourhood during the past 6 months?); and ‘smokevehicle’ (Does anyone smoke cigarettes in a car, taxi or bus while you are riding in it?).

Data gathered from the caregivers includes the variables ‘birthweight’ (How much did the child weigh at birth?); ‘neo-natal’ (Did the child receive any newborn care in an intensive care unit, premature nursery or any other type of special care facility?); and ‘smokepregnant’ (Did the child’s biological mother smoke at any time while she was pregnant with the child?).

Information about the home and environment includes ‘work ‘n wear’ (Is there anyone whose paying job is working around chemicals (such as pesticides, paints) or dust living in the home? / If yes, do they usually wear their work clothes home?); ‘pet’ (Do any pets live in this home? – this was confined to the presence of cats and dogs only); ‘numpeople’ (How many adults (18 years or older) usually live in your home? + How many children (less than 18 years of age) regularly live in your home?); ‘food’ (Which one of the following statements best describes the food eaten by your household? [enough food to eat; sometimes not enough food to eat; often not enough food to eat]); ‘stove’ (What is the primary source of heat for the stove or oven? [paraffin; gas; electricity; wood; coal]); ‘income’ (Which category best describes your total combined income of all members of your household during the last month? [less than R1000; … R10001 and above]); and ‘smokers’ (Does anyone who lives here smoke cigarettes in the home?).

The variable ‘area’ refers to South Durban and North Durban while the ‘asthma severity’ variable is a 4-tiered classification of asthma severity based on criteria provided by the US National Asthma Education Program (NAEPP, 1991) (Table 2.2).

**Table 2.1: Categories and frequencies for all variables**

| VARIABLE | CATEGORIES - COUNT (N=382) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **gender** | male | 163 | female | 219 | | | | |
| **exercise** | <twice weekly | 113 | 2-4 time/week | 135 | >4 times/week | 110 | | |
| **TV** | <1hr a day | 86 | 1 - 3 hr/day | 193 | >3 hrs/day | 78 | | |
| **smokers** | yes | 187 | no | 194 | | | | |
| **breakfast** | daily | 236 | not daily | 121 | | | | |
| **pets** | yes | 114 | no | 264 | | | | |
| **food** | enough | 265 | not enough | 85 | | | | |
| **work 'n wear** | yes | 36 | no | 332 | | | | |
| **smokepregnant** | yes | 35 | no | 328 | | | | |
| **neo-natal** | yes | 50 | no | 318 | | | | |
| **birthweight** | up to 2.5 kgs | 56 | >2.5 kgs | 280 | don't know | 42 | | |
| **fear** | yes | 165 | no | 192 | | | | |
| **violence** | yes | 185 | no | 169 | | | | |
| **weapons** | yes | 160 | no | 194 | | | | |
| **perceived weight** | overweight | 54 | underweight | 35 | correct weight | 267 | | |
| **smokevehicle** | yes | 94 | no | 259 | | | | |
| **stove** | paraffin | 6 | gas | 3 | electric | 308 | no stove | 27 |
| **numpeople** | 1 - 4 people | 124 | 5 - 7 people | 153 | >7 people | 70 | | |
| **age** | 8-9 years | 25 | 10 years | 196 | 11 years | 135 | 12+ years | 26 |
| **income** | Up to R1000 | 79 | R1001 - R4500 | 102 | R4501 - R10000 | 88 | R10001 + | 39 |
| **area** | South Durban | 197 | North Durban | 195 | | | | |
| **asthma severity** | moderate/severe | 27 | mild persistent | 47 | mild intermittent | 76 | no asthma | 232 |

**Table 2.2: Classification of asthma severity**

| | | ASTHMA SEVERITY CLASSIFICATION | | |
|---|---|---|---|---|
| | | Mild intermittent | Mild persistent | Moderate/ severe |
| FEATURE | Symptoms in the day | up to 2 days/week | 3 - 6 days / week | Daily |
| | Night time awakenings with symptoms | up to 2 times / month | 3 - 4 times / month | At least twice a week |
| | Interference with normal activity | None | Minor limitation | Some/ extreme limitation |
| | Symptoms refer to: cough, shortness of breath, wheezing or chest tightness | | | |
| | Patients are assigned to the most severe category in which ANY feature occurs | | | |
| | Adapted from: Guidelines for the Diagnosis and Management of Asthma, NAEPP, Expert Panel Report 3, Pages 305-310. www.nhlbi.gov/guidelines/asthma | | | |

## 2.3 Missingness

Non-response in surveys is a common problem and this data set is no exception. Missingness affects 43.5% of the 382 child records used in this study. It varies across 18 of the 22 variables and amounts to a total of 5.3% of the data items (Figure 2.1). The frequency of non-response per variable is summarized in Figure 2.2.

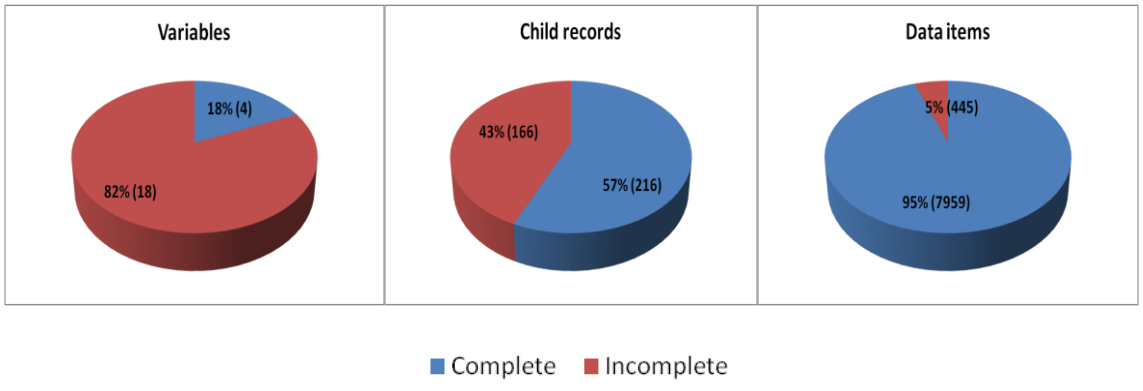



**Figure 2.1: Summary of missing values**

Examining the raw data for each missing item revealed that out of the 382 participating subjects, four did not complete a 'caregiver' survey, there were 23 missing 'child' surveys and 16 missing 'family' surveys'.

Analysis of the missingness present in the variables 'birth weight', 'neo-natal' and 'smokepregnant' from the 'caregiver survey', revealed the following: apart from the four missing surveys, 42 respondents indicated that they did not know the birth weight of the child; and there were 10 non-responses to the question regarding neo-natal care and 15 missing

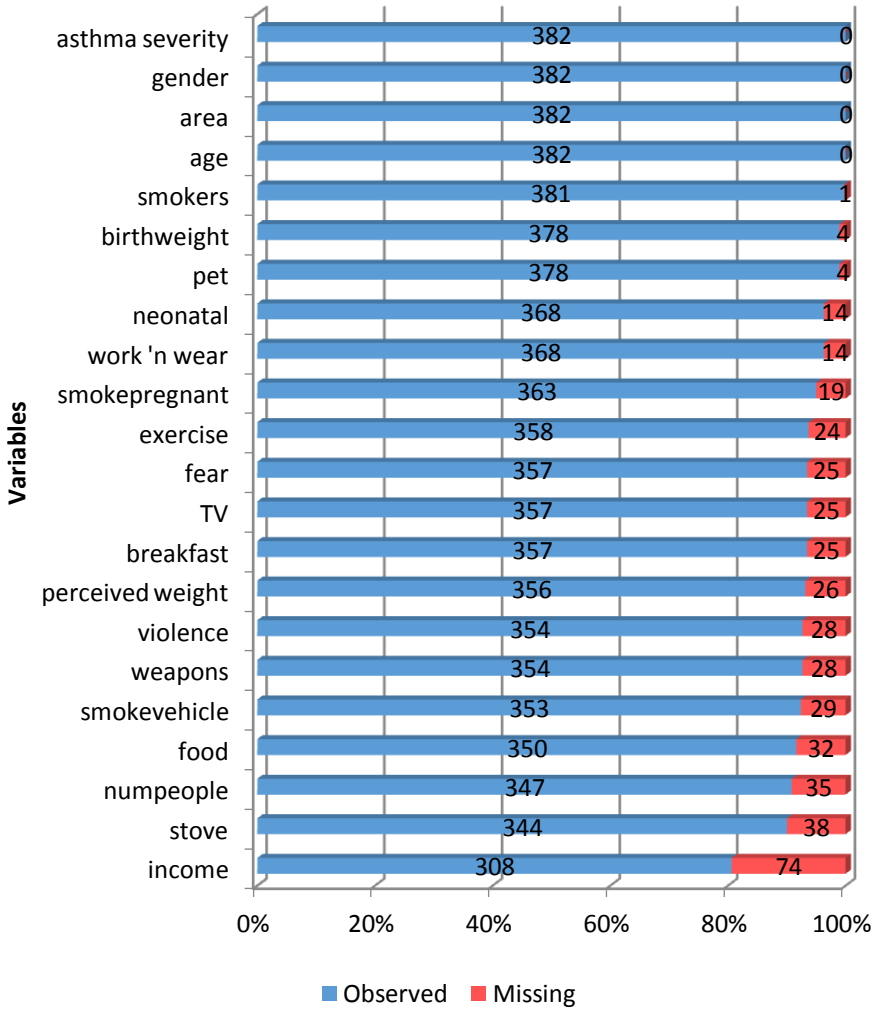



**Figure 2.2: Frequency of missingness for each variable**

items from the question regarding smoking while pregnant. It would seem that in many cases, these missing items were as a result of the respondent to the 'caregiver' survey genuinely not knowing the information asked for since nearly 60% of these respondents were family members, other than the mother, or foster parents.

Information for the 'work 'n wear', 'pet' and 'smokers' variables were obtained from multiple surveys. The few missing items were as a result of either missing surveys, or complete sections, or individual questions, being skipped in the completion of a survey. It seems reasonable that, given the respondent to a survey is not always the head of the household, knowledge of whether a resident in the home works with chemicals and wears his/her work clothes home, would sometimes not be known to the respondent.

The group of socio-economic variables – 'numpeople', 'Income', 'food' and 'stove' – are all from the 'family' survey used to gather information about the home and its inhabitants. These variables have the highest counts of non-response amongst all of the variables. It has been suggested that SES is a sensitive area of inquiry for research studies with high percentages of item non-response (Shavers, 2007). Specifically, 'income' typically has one of the highest rates of item non-response in surveys (Pleis and Cohen, 2007). It has been found that item non-response for household income generally ranges from 21% to 39% (Moore et al., 2000).

Apart from the 16 missing 'family' surveys, there were numerous cases of incomplete sections or individual items not being answered. It is noteworthy that, for the 'income' variable, over 8% of the respondents included in this study completed the full section on income that gathered information about those in the home who receive an income, as well as the source of the income, but failed to respond to the question that asked what their income was.

The remaining incomplete variables are from the 'child' survey. Non-response was largely due to the 23 missing surveys and was supplemented by some individual missing data items and, in a few cases, complete sections being skipped.

From the exploration of the missingness in the raw data, it would appear that much of the missingness can be classified as MCAR. However, especially when one considers the socio-economic variables, it is highly possible that there is some MNAR at play as well.

Chi-square analysis was carried out to ascertain whether the missingness in any of the variables is significantly related to the fully measured variables of 'gender', 'age', 'area' and 'asthma severity'. It was found that missingness in 'birth weight' ($p = 0.032$), 'neo-natal' ($p < .0005$) and 'smokepregnant' ($p=.001$) is related to the 'gender' variable; missingness in 'exercise' ($p = 0.015$), 'TV' ($p = 0.010$), 'fear' ($p = 0.030$), 'breakfast' ($p = 0.010$), 'perceived weight' ($p = 0.002$), 'violence' ($p = 0.025$) and 'smokevehicle' ($p < 0.0005$) is related to 'area'; missingness in 'stove' ($p = 0.042$)is related to 'age' and missingness in 'food' ($p = 0.029$) is related to 'asthma severity'. Thus all of these variables that suffer from missingness could be considered at worst to be MAR. It cannot be ruled out, however, that there exists some MNAR mechanism in the data.

### 2.3.1 Missing value patterns

The missingness values follow a non-monotonic pattern (Figure 2-3). There are 50 individual patterns present in this data, each of which represents a different combination of missing variables. Pattern 1 represents the 216 cases that do not have any missing data. Of those patterns that include missingness, the most common is pattern 35 which includes 39 cases with missingness only on 'income'. This is followed by: pattern 29, with 14 missing data items on 'stove' only; pattern 46 with 12 missing data on 'food', 'numpeople', stove' and 'income'; and pattern 25 with 11 missing data on 'numpeople' only. It is noteworthy to mention that these four variables that suffer from the most missingness are all socio-economic variables from the 'family' survey. One can also easily identify the blocks where complete surveys are missing.

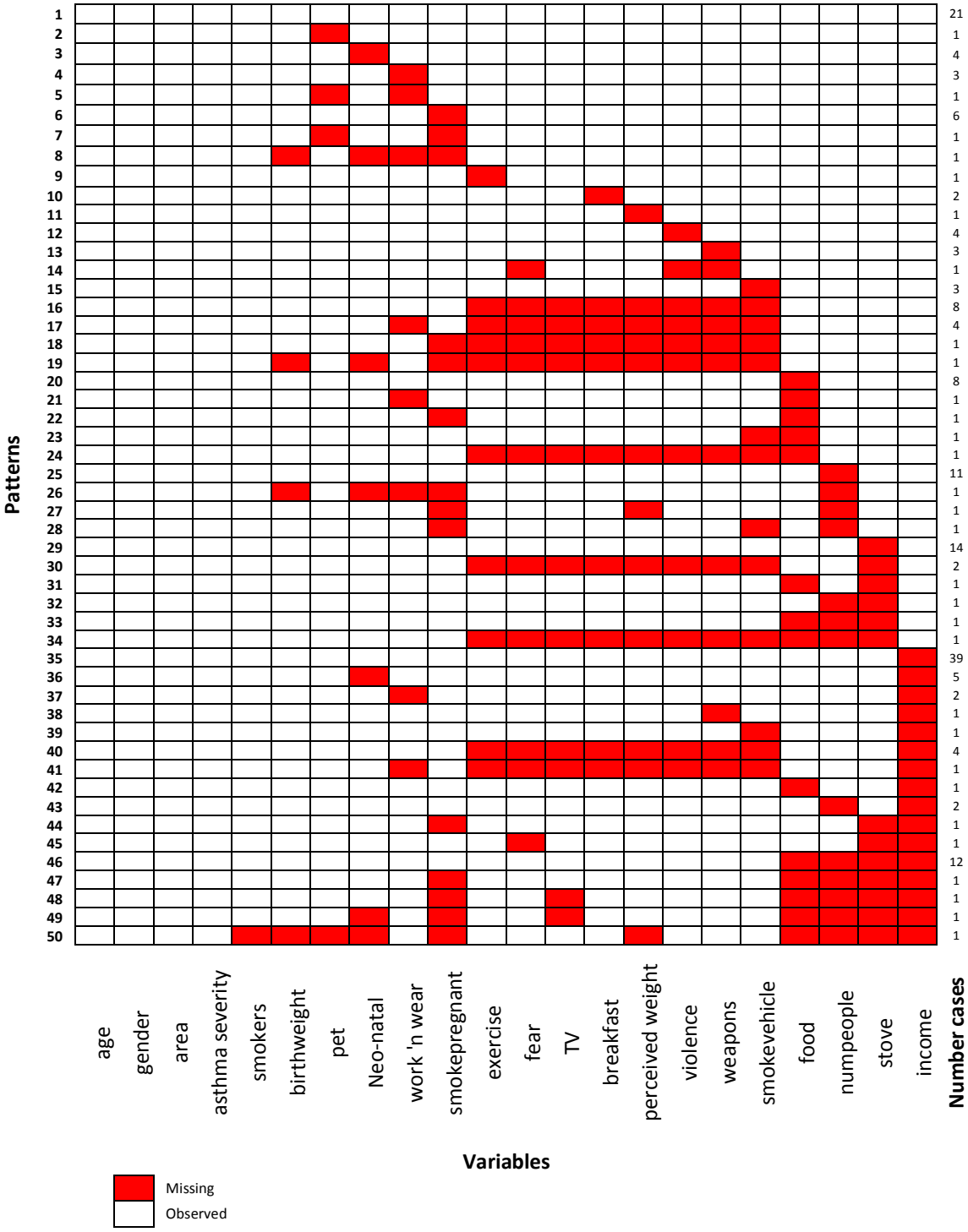



**Figure 2.3: Missing value patterns**

# Chapter 3

# METHODOLOGY

In this chapter a basic description of the theory behind both MI and subset correspondence analysis is presented. In addition, methodologies adopted in their different applications are outlined.

## 3.1 Multiple imputation (MI)

MI is a three-stage process. In the first stage, missing values are imputed. Multiple imputed data sets are thus generated with different imputed values replacing the missing values in each set. The second stage involves the analysis of each of the completed data sets. Parameter estimates and standard errors are retained from each analysis. In the third stage, results from the individual analyses of stage two are combined using what are commonly known as Rubin's rules (Rubin, 2004b).

Although, in the past, it was widely thought that as few as 3 imputed data sets are needed to obtain good results and inferences, new studies have shown that this may, in fact, not be enough (Graham et al., 2007). Studies have shown that there could be an important reduction in statistical power if the number of imputations, m, is small (Graham, 2012). Graham et al. (2007) completed a simulation study on the number of imputations needed to attain maximum power. Their recommendations for the number of imputations, m, as a function of the fraction of missing information are summarized in Table 3.1. On the basis of the percentage of data missing in this study (5.3%), 20 sets of data were imputed.

In this study two different algorithms are considered to carry out the MI – multivariate normal imputation (MVNI) and fully conditional specification (FCS). The MVNI algorithm is a general purpose imputation application that assumes that the data follow a multivariate normal

**Table 3.1: Recommended number of imputations needed for varying fractions of missing data**

| Fraction of missing data | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|---|
| Number of imputations | 20 | 20 | 40 | 100 | >100 |

distribution. FCS is a more flexible approach to MI than MVNI since it is able to handle all types of data including continuous, binary, categorical and ordinal. Given that this study is concerned with categorical data, it would seem that FCS is a better choice for the imputation. However, it has been suggested that results from MVNI may often be sound even if multivariate normality does not hold as in the case of binary and categorical variables (Lee and Carlin, 2010). It is for this reason that both methods were studied and results compared both across methods and against a complete case analysis (Chapter 4).

### 3.1.1 Theoretical background

**Imputation**

MVNI – This imputation algorithm, adopted by the NORM software(Schafer, 1999), assumes the complete data (observed and missing values) follows a multivariate normal distribution. NORM uses a data augmentation (DA) procedure to impute multiple sets of data using parameter estimates obtained from the EM algorithm as starting values.

The EM algorithm for covariances matrices, as applied in MI, calculates sufficient statistics – building blocks of the particular analysis being done – and produces relevant parameters. In this case sufficient statistics are sums, sums of squares and sums of cross products; while relevant parameters are a variance-covariance matrix and vector of means.

The EM algorithm is a two-step iterative procedure that goes back and forth between the E-step and the M-step.

In the E-step, missing values are replaced by scores from a series of regression equations such that each missing variable for a specific case is regressed on the remaining observed variables for that case. Using these observed and imputed values, the sufficient statistics are calculated.

In the M-step, ML estimates of the mean vector and covariance matrix are obtained using the sufficient statistics calculated at the previous E-step. The resulting covariance matrix and regression coefficients from the M-step are then used to derive new estimates of the missing values at the next E-step and the process begins again.

The algorithm repeatedly cycles through these two steps until the difference between covariance matrices in subsequent M-steps satisfies some convergence criterion. The variance-covariance matrix and vector of means thus produced are ML estimates of these quantities.

The data augmentation that follows EM is also a two-step process. In the first step, DA randomly imputes the missing data using the assumed values of the parameters. In the second step, new parameter estimates are drawn from a Bayesian posterior distribution based on the observed and imputed data. The repetition of these two steps results in a Markov chain. DA converges when the distribution of parameter estimates stabilizes. Research has shown that DA nearly always converges in fewer cycles than does EM (Schafer and Olsen, 1998). This enables one to estimate the cycle length, k, of DA as being any number at least as large as the number of iterations needed for EM to converge.

In order to impute m sets of data, DA is run for N = mk iterations and the data set at the end of every $k^{th}$ cycle is saved.

When the data contain categorical variables, some adjustments are necessary both before and after imputation. Before imputation, dummy coding is applied to all the categorical variables and interaction product terms with more than two categories. After imputation, sensible rounding (Allison, 2002) is used on these variables to prepare the data for analysis.

FCS – FCS, also termed ‘chained equations’, is the MI algorithm adopted by SPSS (SPSS inc.). This is a more flexible approach to imputation in that it is designed to handle different types of

variables (continuous, binary, categorical, ordinal) and does not assume multivariate normality of the data (Lee and Carlin, 2010).

In practice, FCS involves running a series of regression models such that each variable with missing data is regressed on the other variables in the data set according to its distribution. So, for example, categorical variables will be modelled using logistic regression and continuous variables will be modelled using linear regression.

Imputation by FCS, as applied in SPSS, is also an iterative process that starts by imputing every missing value with random draws from the distribution of the non-missing values. Continuous variables are replaced with draws from a normal distribution and categorical variables are replaced with draws from a multinomial distribution. Azur et al. (2011) refer to these replacements as 'place holders'.

Each iteration involves the following steps:

- Set the 'place holders' of one variable that suffers from missing values back to missing
- Set up a regression equation, according to the distribution of the variable, with the observed values as the dependent variable and the other variables as independent variables
- Replace the missing values from this variable with predictions from the regression equation
- Repeat these steps for each variable that has missing values.

This forms one iteration of the process. At each iteration the imputed values are updated. This process is repeated for a specified number of iterations, n, after which the data set is retained as one complete imputed data set. The number of iterations, n, chosen so that the parameters from the regression models have stabilized, is generally about ten (Raghunathan et al., 2002). This entire process is repeated until the required number, m, of imputed data sets is generated.

**Ordinal Regression**

Because the data in this study is primarily categorical with an ordinal dependent variable – asthma severity, ordinal regression – an extension of logistic regression – was chosen to analyse the m imputed data sets.

Suppose the dependent variable, $Y$, has $J$ levels ordered in increasing order of magnitude. Let the probability that $Y$ takes on the value at level $j$ be defined as $\Pi_j = P(Y = j)$.

For a given set of $p$ independent variables $\boldsymbol{x} = (x_1, x_2, \dots x_p)$, the generalised logistic model takes the form of

$$logit\,(\pi_j) = ln\left(\frac{\pi_j}{1-\pi_j}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots . + \beta_p x_p \qquad \mathbf{(3.1)}$$

In ordinal regression, this model is modified to reflect the ordinal characteristics of the dependent variable by using one of a selection of link functions. In this study, the cumulative logit link function is applied with an additional constraint imposed on the logit coefficients $(\beta_1, \dots, \beta_p)$, such that they are the same across all $J - 1$ logits. This results in the following ordinal logit model:

$$logit\left(\pi\left(Y \le j \middle| x_1, x_2, \dots, x_p\right)\right)$$

$$= \ln\left(\frac{\pi\left(Y \le j \middle| x_1, x_2, \dots, x_p\right)}{\pi\left(Y > j \middle| x_1, x_2, \dots, x_p\right)}\right)$$

$$= \beta_{0j} + \left(-\beta_1 x_1 - \beta_2 x_2 - \dots - \beta_p x_p\right) \qquad \mathbf{(3.2)}$$

This model is known as the proportional odds model with a cumulative logit link.

Note that the negative signs before the logit coefficients in equation (**3.2**) are to facilitate interpretation so that larger coefficients indicate an association with higher levels of the dependent variable.

It can be seen that, while each of the $J-1$ cumulative logits has a unique intercept value, $\beta_{0j}$ – called the threshold value, the values of the logit coefficients are the same across all logits. This imposed constraint as adopted by many applications, including SPSS, is tested using the 'parallel lines test'.

When interpreting results from ordinal regression, it can be useful to examine the odds ratios. By calculating the odds ratio, $e^{\beta}$, for any given category of an independent variable, it is possible to determine the odds of scoring a higher value on the dependent variable, relative to the reference category of that independent variable.

### 3.1.2 Methodologies adopted in the applications of MI

***Identification of interactions***

Prior to the imputation stage it was necessary to identify interactions present in the data to be included at the analysis stage to ensure that missingness was as close to MAR as possible. This was achieved by analysing a single data set imputed from EM parameters. This single imputed data set was generated using the EM step in the NORM software (Schafer, 1999). The default settings for maximum iterations (1000) and for convergence (.0001) were used and the standard maximum likelihood estimates were requested. All other default options regarding output and file names were retained. The strategies used in the identification of interactions and main effects are detailed in Chapter 4.

***Imputation of complete data sets***

For both imputation methods, all variables, the outcome variable (asthma severity) and the identified interactions were included in the imputation model.

**MVNI**

MVNI was performed using the stand-alone NORM software (Schafer, 1999). There are many steps to the process of running NORM – some are standard and others depend on the specific application. Details of the general steps common to all applications can be found in Graham (2012). Those steps specific to this study are outlined below.

Preparing the data:

Because MVNI assumes normality of the data, adjustments to the data were needed prior to imputation. No recoding was necessary for the continuous variable (age) or any of the dichotomous categorical variables before imputation. However, the categorical variables with more than two levels (exercise; TV watching; income; perceived weight; stove type; number of people; birth weight and asthma severity) and the 10 identified interactions were dummy coded such that a variable with *p* levels was represented by *p-1* dummy variables.

All missing values were set to -999.

Variables:

In order to ensure that all variables were of the required scale type (continuous, dichotomous and categorical) for analysis, some adjustments were necessary to the imputed values. For all the dichotomous variables, selecting the word 'integer' in the rounding column under the 'variables' tab was sufficient. For the interactions and categorical variables with more than two levels, automatic rounding of the dummy variable to 0 (absent) or 1 (present) for each of the levels was not always possible since it sometimes happened that more than one category rounded to one. In this case, 'sensible' rounding (Allison, 2002) was performed which entailed selecting the category with the largest imputed value and rounding it to one with all the other levels to being rounded to zero.

EM Algorithm:

This step was used to obtain starting values for the imputation process. The same settings were used as for the application of the EM algorithm in the identification of the interactions.

Data Augmentation using NORM:

Under the *Series* button, 'Save all parameters at every kth cycle' was selected and k = 1 specified.
Under the *Imputation* button, 'Impute at every kth iteration' was selected and k = 36 was specified. The EM process took 36 iterations to converge, hence the choice for k.
Under the *Computing* button, the number of iterations was specified as 36 x 20 = 720.

**FCS**

FCS was carried out using the MCMC algorithm available on SPSS (version 17) (SPSS inc.). The steps followed and specifications adopted follow:

Preparing the data:

All variables were defined in the variable view tab including the type of variable
Incomplete variables were defined as nominal or scale prior to imputation

Variables tab:

Variables for the imputation model were selected.
The number of imputed data sets was set to 20.
File name for output data was specified.

Method tab:

The fully conditional specification (MCMC) option was selected.
Maximum iterations was specified as 10 (the default option).

Output tab:

Imputation model was selected.
Create iteration history was selected.

All categorical variables with missing data were imputed using logistic regression. The imputed data sets were stacked into a single file with a variable named IMPUTATION_ to differentiate between the data sets.

***Analysis of the multiple imputed data sets***

For each of the imputation methods the 20 imputed data sets were analysed. Since the dependent variable, asthma severity, is an ordered variable, the analysis tool used for this application was ordinal regression. The logit link function was applied.

In applying ordinal regression with the logit link, it is assumed that the relationship between the independent variables and the log of the odds (logit) of a dependent variable is the same, in a statistical sense, for all dependent variable categories. This means that the regression coefficients for the independent variables are the same across all logits. This important assumption is tested using the test of parallel lines. This test has been described as 'anti-conservative' in that 'it nearly always results in the rejection of the proportional odds assumption' (O'Connell, 2006). Some reasons for failure of the test are: a large number of explanatory variables (Brant, 1990) ; large sample size (Allison, 1999) ; or the presence of a continuous explanatory variable (Allison, 1999). Failure of this test can compromise the results. In the initial application on the data used in this study, the test of parallel lines failed. Several solutions were considered to deal with the problem. While the reduction in the number of explanatory variables or the exclusion of the continuous variable (age), or its recoding into a categorical variable, might have solved the problem, all the variables were considered

important to the study and so this option was not considered. Another possibility was to apply multinomial logistic regression instead of ordinal regression. This, however, while having less stringent assumptions, would result in a loss of power and was, therefore, not considered further. Another solution, suggested by Garson (2008) was to combine categories until parallelism is achieved. It was found that by combining the two lowest categories of asthma severity – no asthma and mild intermittent asthma – the problem was solved. Hence, for the applications in this study, the dependent variable (asthma severity) has only three categories – moderate to severe, mild persistent and mild intermittent or no asthma.

***Combining results from multiple imputation and analysis***

In order to obtain statistical inference from MI, parameter estimates were combined following Rubin's rules (Rubin, 2004b). The two quantities that were dealt with in this manner were the point estimate of the regression coefficients and the standard errors.

The point estimate for each regression coefficient was calculated as the arithmetic mean of the regression coefficients across the 20 imputed data sets.

The standard error for each regression coefficient is broken down into two parts – within-imputation variance and between-imputation variance. Within-imputation variance (*W*) reflects normal sampling variability found in all analyses. It was calculated as the average of the 20 squared standard error (SE) values resulting from the analyses of the 20 imputed data sets. Between-imputation variance (*B*) is a measure of the uncertainty or added variability due to the missing data. It was calculated as the sample variance of the regression coefficient across the 20 imputed data sets. The total variance (*T*) is the weighted sum of these two afore-mentioned variances and was calculated as:

$$T = W + \left(1 + \frac{1}{20}\right) B \qquad (\mathbf{3.3})$$

The standard error used in the inference following MI was the square root of *T*.

## 3.2 Subset correspondence analysis

In contrast to MI, and as yet not popularly adopted as a tool to manage missing data, is subset correspondence analysis (s-CA) - a variant of correspondence analysis (CA). The use of this relatively new method in the management of missing data forms a major part of this study.

Correspondence analysis, as we know it today, is a graphical technique used in many disciplines to study relationships between the rows and columns of a matrix of non-negative numbers. A set of data in multi-dimensional space can be reduced to a lower dimensional space such that associations between variables are easily identified. According to Greenacre (1984), the algebra of CA can be traced back to the 1930's when H. O. Hartley (also known by his original German name Hirschfeld) published an article outlining the mathematical formulation of the association between two quantitative variables in a two-way contingency table (Hirschfeld, 1935). Over the following 30 years, several researchers independently developed the same theory but in different contexts. These included Fisher (1940), Horst (1935), Guttman (1941), Hayashi (1950) and Richardson and Kuder (1933) . It was not until the early 1960's that the geometric form known as 'correspondence analysis' was first published. In this context, the word 'correspondence' refers to 'associations'. Jean-Paul Benzécri, the French researcher responsible for this development, along with a group of data analysts, worked extensively in the area of descriptive multivariate techniques including CA and developed their own philosophy on data analysis. They believed that the data and how they are described is what is important and not the model that one may think the data fit. The statistical techniques developed by the group contained rigorous algebraic notation and were based on geometry which resulted in the graphical displays commonly associated with correspondence analysis.

While CA is similar to several other techniques used to perform multivariate analysis, it is not the same as any of them. CA 'derives sets of multidimensional 'scores' with a well-defined and intentional geometric interpretation' (Greenacre, 1984). Over time, several variants of the original 'simple correspondence analysis' have been developed. These include multiple correspondence analysis (MCA); joint correspondence analysis (JCA) and subset correspondence analysis (s-CA). s-CA involves the application of CA to a subset of the data. This

variant has facilitated the analysis of the subset of the measured data, thus excluding the missing data, without the loss of any information. CA as applied to a subset is implemented as the function ca in R (RDevelopment_CORE_TEAM, 2006).

### 3.2.1 Theoretical background

Correspondence Analysis (CA) is an exploratory multivariate technique applied to any matrix of non-negative numbers in order to identify associations present in the data. In CA, the rows and columns of the matrix are represented by two separate clouds of points in multi-dimensional space. CA finds respective subspaces of low dimension that optimally contain these clouds of points. The principal axes are chosen such that the inertia of the clouds of points is maximised. The inertia of these clouds can be considered as a measure of dispersion or spread of the points taking into account both distance and attributed weights, called masses. CA thus provides a visual interpretation of the relative positions of both clouds in a common subspace of low dimension. Interpretation of the axes can be achieved by examining the decomposition of the inertia of each cloud of points along the principal axes and amongst the points themselves (Greenacre, 1984). By studying the contributions that the points make to the principal axes and the contributions that the axes make to the inertia of the points, those points that are well defined in a plane can be identified. Using these points, it is usually possible to assign 'meanings' to the principal axes. Graphically, if the angle between this point vector and the axis is small, then the point is highly correlated with the principal axis. The distance between two points (either two row points or two column points) is said to be a 'weak' approximation of the chi-square distance between the vectors of relative frequencies of the points (Greenacre, 1978) . One can get an idea of how close two points are by examining the angle the point vectors make with each other. The smaller the angle, the closer they are related. The interpretation of the graphical display is primarily done on the basis of where a point, or group of points, is positioned relative to the axes in the plane.

The variables used in the calculation of the subspace are called active variables. It is possible to examine the position of additional variables, called supplementary variables, relative to this space. These variables play no part in the determination of the principal axes and the optimal subspace but are projected onto an existing subspace. Relationships between these variables, both active and supplementary, and the principal axes can be explored (Greenacre, 1984,

Greenacre and Blasius, 2006). In practice, the associations of the active variables are displayed and then the supplementary variables are related *a posteriori* to these associations (Greenacre and Pardo, 2006b).

In the same way that CA is applied to a full set of data, s-CA is applied to a subset of the data. An appealing feature of s-CA is that, as the full data matrix, ***N***, can be partitioned into a number of separate non-overlapping and all-inclusive matrices, so is the inertia of the full matrix equal to the sum of the inertias of the separate matrices (Greenacre and Pardo, 2006a).

So, if $\boldsymbol{N} = [\boldsymbol{N}_1 : \boldsymbol{N}_2 : \boldsymbol{N}_3]$ , it follows that the inertia of ***N***, In(***N***), follows the rule

$$\text{In}(\boldsymbol{N}) = \text{In}(\boldsymbol{N}_1) + \text{In}(\boldsymbol{N}_2) + \text{In}(\boldsymbol{N}_3) \qquad \textbf{(3.4)}$$

Thus one is able to see how much of the total variation in the data is accounted for in each sub-matrix.

A description and basic calculations of s-CA as applied to a matrix ***N***, in the form of a contingency table, is presented below. Further details can be found in Greenacre (1984), Greenacre and Pardo (2006a) and Greenacre (1992).

From the matrix ***N*** of non-negative numbers $n_{ij},\ i = 1, \dots, I \ and \ j = 1, \dots, J,$ the correspondence matrix, ***P***, is formed by dividing each element of ***N*** by its grand total such that

$$p_{ij} = \frac{n_{ij}}{\sum_i \sum_j n_{ij}} \qquad \textbf{(3.5)}$$

with row and column sums, ***r*** and ***c***, of P defined by

$$r_i = \sum_j p_{ij} \ \ , i = 1, \dots, I \qquad \text{and}$$

$$c_j = \sum_i p_{ij} \ \ , j = 1, \dots, J \qquad (\mathbf{3.6})$$

The elements of ***P*** can be thought of as the probability density of the cells of the matrix and the vectors of row and column sums of ***P***, as marginal densities. The elements of **r** and **c,** termed masses, are a measure of the relative importance of each row and column point. They are represented in diagonal matrices as $\boldsymbol{D_r}$ and $\boldsymbol{D_c}$ respectively. By dividing each element of a row (column) by its respective row (column) sum, we form a vector of relative frequencies which is called a row (column) profile. These profiles define the two clouds of points, one for rows and one for columns, in multi-dimensional weighted Euclidean space. The dimension weights for the row and column clouds are defined by the inverse of the elements of **c** ($\boldsymbol{D}_{\mathbf{c}}^{-1}$) and **r** ($\boldsymbol{D}_{\mathbf{r}}^{-1}$) respectively.

Under the assumption that the rows and columns of ***P*** are independent, the expected value of cell (i,j) of ***P*** is the product of the masses, $r_i c_j$. Centring and normalising the correspondence matrix results in a matrix of standardised residuals ***S*** such that

$$s_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} \qquad (\mathbf{3.7})$$

The sum of squared elements of ***S*** is a measure of the total variation in the data and is termed total inertia.

It is at this stage that the CA process is 'interrupted' to implement the 'adjustment' needed for s-CA.

From the matrix, ***S,*** of standardized residuals, select those rows and columns that make up the subset of variables/categories chosen to be included in further analysis. Let this matrix be ***S*.*** It is important to note that marginal densities, **r** and **c**, for the full matrix are retained for all future calculations (Greenacre and Pardo, 2006a).

The objective of CA and its variants, including s-CA, is to identify low dimensional subspaces of the row and column clouds which are closest to the points in terms of weighted sum of squared distances. This is achieved by performing a singular value decomposition (SVD) on ***S*.*** In other words, $\boldsymbol{S^*} = \boldsymbol{U\Delta V^T}$, where ***U*** and ***V*** are the left and right singular vectors, respectively,

and ***Δ*** is a diagonal matrix of singular values in decreasing order of magnitude. The principal axes of the row and column clouds are defined, respectively, by the ***K**** left and right singular vectors corresponding to the ***K**** largest singular values.

From the result of the SVD, we are able to define the principal co-ordinates of the points, i.e. co-ordinates with respect to their principal axes.

The principal co-ordinates of row $i$ and column $j$ on dimension $k$ are defined, respectively, as

$$f_{ik} = \frac{u_{ik}\lambda_k}{\sqrt{r_i}} \qquad \text{and}$$

$$g_{jk} = \frac{v_{jk}\lambda_k}{\sqrt{c_j}} \qquad (\mathbf{3.8})$$

It is these co-ordinates that are used to produce the graphical displays of the points.

The amount of inertia explained by each principal axis is given by the square of the corresponding singular value.

### 3.2.2 Methodologies adopted in the applications of s-CA

***Data preparation and analysis***

CA can be applied to any data providing they are non-negative and categorical. The continuous variable – age – was therefore recoded into four categories (8-9 years; 10 years; 11 years and 12+ years) for all applications.

For each variable with non-response, a separate missing category was introduced. Raw data were presented in the form of a contingency table, where the asthma categories were cross tabulated with the other variable categories. The rows represent the asthma categories and the columns represent the categories, including the missing categories, of the other selected variables.

To specify the subset for analysis, all column categories representing the missing data were labelled for exclusion. Where necessary, columns for treatment as supplementary columns were labelled as such. Principal co-ordinates were used to plot the variables in all the graphical displays.

Where applicable, interactions were added to the data by forming dummy variables (yes/no) for all combinations of the cross variable categories. These dummy variables are treated as additional variable categories.

Where applicable, individual variables involved in interactions were tagged to be treated as supplementary variables.

***Interpretation of output***

Graphical plots (as seen in figures 5.1, 6.1, 6.6 and 7.1) as well as numerical output showing both the decomposition of inertia (as seen in Tables 5.1, 6.2 and 7.2) and total inertia were used in the interpretation of the results.

With reference to the plots, by joining each point to the origin with an imaginary line, it was possible to identify the strength of association between the points. The smaller the angle they make with each other, the closer is the association. This concept was also applied to the association of points with the axes. In addition, variables that are not well represented in the subspace are situated near the origin and do not add appreciably to the interpretation of the display.

The total inertia, a measure of the variability in the data, and its decomposition along the axes and among the points were all used to aid in the interpretation of the results. By examining the percentage of the total inertia that is represented on each axis, it was possible to identify the relative importance of the axes and the amount of variability in the data that they represented.

It was also possible to give ‘meanings’ to the axes by examining the absolute contributions (labelled CTR in the output) that the points make to the inertia of the axes. A row/column

point is deemed important to the orientation of an axis, if its CTR value exceeds ‘1000/the number of row/column points’ respectively.

In the same way, by examining the relative contributions (labelled COR in the output) for each point, it was possible to identify the axis which best represents the point. These values are a measure of how close a point lies to each of the axes and are independent of its mass or distance from the origin. High values of COR indicate that the axis contributes highly to the point’s inertia; the angle the point makes with the axis will be small and the point is said to ‘correlate’ with the axis. Points with extremely high COR values are positioned nearly on the axis; this indicates that there is very little error in its location on the display.

The sum of the COR values across the dimensions is represented in the QLT column. This was used to indicate the quality of representation of the points in the subspace of chosen dimensionality. Values have been scaled so that, across all possible dimensions, QLT equals 1000.

# Chapter 4

# IDENTIFYING INTERACTIONS FOR MULTIPLE IMPUTATION IN THE PRESENCE OF MISSING DATA

MI is a reliable tool to deal with missing data and is becoming increasingly popular in biostatistics. However, building a model with interactions that are not specified *a priori*, in the presence of missing data, presents a challenge. On the one hand, the interactions are needed to impute the data; while on the other hand, the data are needed to identify the interactions. In this chapter, two strategies are investigated in which model development, with interactions, is achieved using a single data set generated from the Expectation Maximization (EM) algorithm (4.2). Imputation using both the FCS approach and the MVNI approach is carried out and results are compared (4.4). These results are further compared to a complete case analysis in which only those child records with a full set of measured data were included. The theory of these imputation algorithms was presented in Chapter 3.

## 4.1 Background

MI is successfully applied to data that are MAR and yields unbiased results with accurate estimates for the standard errors (Donders et al., 2006). Unfortunately, the missingness mechanism is not usually fully known and is often a combination of more than one mechanism. However, by ensuring that the imputation model is more general than the analysis model, MI will usually produce sound results (Collins et al., 2001, Graham et al., 1997, Graham, 2012, Schafer and Olsen, 1998). This is achieved by including, in the imputation model, variables that are related to the incomplete variables as well as those related to their missingness; the outcome variable; and all interactions that will be examined in the analysis.

Rubin (1996) suggests that the need to include all possibly relevant predictors in the imputation model is demanding in practice. If interactions are selected *a priori*, it is a

straightforward exercise to include them in the imputation model (Graham, 2012). If, on the other hand, the relevant interactions have not been identified beforehand, then ideally all possible interactions should be included in the imputation model. This is neither practical nor, in some cases, possible (Schafer, 1997, Stuart et al., 2009), particularly when the number of variables is large. While model development with MI has been documented (Stuart et al., 2009, Vergouwe et al., White et al., 2011, Wood et al., 2008), none of these studies addresses the issue of how to include, in the imputation model, interactions that are not known *a priori*.

## 4.2 Model building for imputations with interactions

In order to ensure that the imputation model is at least as complex as the analysis model, and that the assumption of MAR is plausible, it is necessary to include the outcome variable and all possible likely predictors for the analysis model, in the imputation model. The selection of the interaction terms presents difficulties (White et al., 2011, Wood et al., 2008) . Comparable to the suggestion made by White et al. (2011), a single complete set of data using the EM algorithm for covariance matrices is generated. The EM algorithm is an iterative procedure that can be used to create a complete data set in which all missing values are replaced by maximum likelihood (ML) values that are asymptotically unbiased. The process starts by replacing each missing value with an estimate calculated from a regression equation in which all the other variables are predictors. Once all the missing values have been replaced, a variance covariance matrix and a vector of means from the completed data are calculated. New regression equations are then formed to predict a new set of estimates for the missing values. This process is repeated until the variances, covariances and means converge, thus producing ML estimates of the parameters. The complete data set generated from this process is then used for model development and the identification of interactions.

To develop the best model given the large number of variables available, the following three-stage process is followed: Firstly, all variables are purposefully selected as main effects. Secondly, in developing the full model, interactions are chosen one at a time in a stepwise manner such that the interaction that makes the biggest significant improvement to the fit is added to the model. For this process a cut-off *p*-value of 0.05 is used. Thirdly, when no further improvement to the fit is possible, backward elimination is carried out to find the smallest

model that is as good as the full model. Here a *p*-value of 0.10 is used for the stopping criterion.

In the setting of the MI process, two possible strategies that can be applied to carry out the model development process are suggested. These are applied to both the MVNI and the FCS approaches to MI.

*Strategy 1* (S1). All three stages of the model development process – the selection of main effects, identification of interactions as well as the backward elimination – are performed on the initial data set generated by the EM parameters. The variables and interactions identified by this process are incorporated into the imputation model. Interactions are treated differently, depending on which imputation method is used.

For MVNI as implemented in the NORM software, interactions with p categories are treated as categorical variables and coded into p-1 dummy variables before being added to the raw incomplete data. By way of an example: an interaction between gender (male/female) and smoking (yes/no) is broken down into separate categories – male/yes, male/no, female/yes and female/no – and binary coding (present/absent) is applied to the first three categories.

For FCS, the interaction is coded according to the possible categories. So, in the example above, male/yes = 1, male/no = 2, female/yes = 3 and female/no = 4.

The interactions as coded in the two scenarios above are merely treated as additional variables. This has been referred to as the 'transform-then-impute' method of dealing with interactions and, in a regression model that includes interactions, has been shown to yield good regression estimates, even though the imputed values are inconsistent with one another. In contrast to this is the 'impute-then-transform' method, also known as passive imputation, which yields plausible-looking imputed values but biased regression estimates(Von Hippel, 2009).

This imputation model is then used to produce the m sets of imputed data. These are analysed individually and the results are combined using Rubin's rules (Rubin, 2004b).

*Strategy 2* (S2). Using the initial EM generated data set, the first two stages of the model development process are completed: selection of main effects and identification of interactions. These are then incorporated into the imputation model as before and m sets of imputed data are produced. Analysis, followed by the third stage of model development (backward elimination), is then applied to each of these data sets. The final selection of variables for the model includes those that are selected in at least 50% of the individual data sets. In the event that no variables satisfy the selection criterion, the condition can be relaxed to a lower percentage. Once these variables are established, analysis is carried out on each data set and the results are combined.

## 4.3 Analysis procedures

Given that the outcome variable, asthma severity, is an ordinal measure, the chosen method of analysis for this data is ordinal regression. The three categories of the outcome variable are 'none/mild intermittent asthma'; 'mild persistent asthma' and 'moderate/severe asthma'. For all the analyses, logit is the chosen link function.

In addition to the analysis of the imputed data, a complete case analysis is carried out for comparative purposes. All main effects and interactions that are defined in stages 1 and 2 of the model building process are used with the complete case analysis and then backward elimination is applied to reduce the model.

## 4.4 Results

### 4.4.1 Model building

***Imputed data -MVNI***

The two different strategies suggested for building the model using the imputed data resulted in the identical set of variables and interactions being identified. In each case 17 main effects and 10 interactions were included in the final model (Table 4.1). While fewer than half of the main effects were significant, the interactions in which these variables were involved were largely significant. Main effects dropped from the model include birth weight, perceived

weight, weapons and stove type. However, these were left in the imputation model as they were shown to be associated with other variables and/or their missingness.

***Imputed data -FCS***

Model development following strategy 1 resulted in the identical model as identified when applying MVNI imputation. The set of significant variables from the two analyses were, however, not the same. Two main effects and three interactions differed in their significance. With strategy 2, the variable 'smoke while pregnant' and its interaction with 'area' did not make the cut to be included in the model. These two variables were significant in only 9 of the 20 individual analyses, whereas, they were significant in 10 of the 20 analyses when MVNI imputation was applied.

***Complete case analysis***

The complete case analysis was based on 216 complete cases, representing 56.5% of the total available cases. The final model contained 16 main effects and 7 interactions (Table 4.1).

The main effects selected with the complete case data compared to those selected with the imputed data differed slightly. 'Perceived weight' and 'weapons' are the only variables that are in the complete case model but not in the imputed data model. Three of the 10 interactions and three of the main effects from the imputed data models were not retained in the complete case model. The models from the imputed data contained more variables than the complete case model.

### 4.4.2 Analysis

Results of the three different analyses of the imputed data (Table 4.1) were, in general, very similar. The size and direction of association between asthma severity and all the predictor variables, as well as the standard errors (SE's) of the estimated coefficients were consistent across both types of imputation as well as for both model building strategies. Even though some differences in the significance of certain predictors did occur, in all cases the p-values showing significance of these predictors were only marginally different from the 5% cut-off value.

**Table 4.1: Estimated coefficients (EST) and standard errors (SE) for the predictors selected in the different analyses**

| Predictor | Reference Category | Category | CC(N = 216) | | MVNI (N = 382) | | FCS1(N = 382) | | FCS2(N = 382) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | EST | SE | EST | SE | EST | SE | EST | SE |
| Gender | Female | Male | -0.441 | 0.674 | 0.129 | 0.398 | 0.030 | 0.391 | 0.017 | 0.390 |
| Neo-natal care | No | Yes | 2.484* | 0.723 | 1.103* | 0.444 | 1.112* | 0.450 | 1.085* | 0.446 |
| Fear | No | Yes | -1.169 | 0.649 | -0.958* | 0.431 | -1.009 * | 0.451 | -1.073 * | 0.444 |
| Smoked while pregnant | No | Yes | 4.256* | 1.237 | 1.019 | 0.736 | 0.885 | 0.693 | 0 | |
| Smokers in home | No | Yes | 0.939 | 0.537 | 0.742* | 0.352 | 0.761* | 0.341 | 0.801* | 0.335 |
| Smoke in vehicles | No | Yes | -2.584 * | 0.921 | -0.253 | 1.068 | -0.308 | 1.011 | -0.323 | 1.015 |
| Exercise | >4 times a week | Up to once a week | 2.805* | 1.227 | 0.892 | 0.761 | 0.692 | 0.756 | 0.624 | 0.731 |
| | | 2 – 4 times a week | 3.313* | 1.229 | 1.039 | 0.717 | 0.936 | 0.718 | 0.738 | 0.680 |
| TV watching | >3 hours a day | Up to 1 hour a day | -0.566 | 0.854 | 0.399 | 0.684 | 0.327 | 0.669 | 0.346 | 0.657 |
| | | 1 – 3 hours a day | 0.304 | 0.769 | 0.641 | 0.639 | 0.525 | 0.630 | 0.569 | 0.618 |
| Number people in home | 8+ | 1 - 4 | 0 | | 1.084 | 0.554 | 1.060* | 0.539 | 1.101* | 0.526 |
| | | 5 - 7 | 0 | | 0.226 | 0.552 | 0.254 | 0.551 | 0.250 | 0.540 |
| Income | R100001+ | up to R1000 | 2.840* | 1.257 | 0.695 | 0.8 | 0.787 | 0.789 | 0.823 | 0.778 |
| | | R1001 – R4500 | 1.285 | 1.203 | 0.209 | 0.797 | 0.489 | 0.754 | 0.431 | 0.754 |
| | | R4501 – R10000 | 1.933 | 1.17 | 1.428 | 0.783 | 1.401* | 0.692 | 1.356 | 0.692 |
| Food availability | Enough | Not always enough | -0.575 | 0.64 | 0.604 | 0.503 | 0.665 | 0.464 | 0.677 | 0.455 |
| Perceived weight | Correct weight | Overweight | -0.230 | 0.743 | 0 | | 0 | | 0 | |
| | | Underweight | 2.369* | 0.97 | 0 | | 0 | | 0 | |
| Work'nWear | No | Yes | 0 | | -0.635 | 0.626 | -0.543 | 0.629 | -0.478 | 0.622 |
| Pets ever | No | Yes | -3.770 * | 0.994 | -1.658* | 0.501 | -1.483 * | 0.503 | -1.413 * | 0.467 |
| Area | North Durban | South Durban | 6.278* | 1.461 | 2.042* | 0.76 | 1.948* | 0.737 | 1.597* | 0.671 |
| Breakfast habits | Daily | Not daily | -4.098 | 3.04 | -0.492 | 1.512 | -0.234 | 1.548 | -0.110 | 1.518 |
| Violence | No | Yes | 0 | | -0.817* | 0.382 | -0.741 * | 0.377 | -0.715 | 0.373 |
| Weapons | No | Yes | -1.147 * | 0.555 | 0 | | 0 | | 0 | |
| Age | | | -1.068 * | 0.438 | -0.79* | 0.254 | -0.833 * | 0.268 | -0.834 * | 0.265 |

**Table 4.1: Estimated coefficients (EST) and standard errors (SE) for the predictors selected in the different analyses (continued)**

| Predictor | Reference Category | Category | CC(N = 216) | | MVNI (N = 382) | | FCS1(N = 382) | | FCS2(N = 382) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | EST | SE | EST | SE | EST | SE | EST | SE |
| Fear*Breakfast | No/daily | Yes/not daily | 2.635* | 1.219 | 2.047* | 0.866 | 2.123* | 0.916 | 2.185* | 0.911 |
| Gender*SmokeVehicle | Female/No | Male/yes | 5.092* | 1.342 | 2.535* | 1.034 | 2.431* | 0.977 | 2.464* | 0.971 |
| SmokeVehicle*TV | No/>3 hrs | Yes/up to 1 hr | 0 | | 0.891 | 1.298 | 0.675 | 1.265 | 0.722 | 1.250 |
| | | Yes/1 – 3 hrs | 0 | | -2.184* | 1.085 | -1.975 | 1.034 | -2.002 | 1.037 |
| Food*Age | enough/ | Not always enough/ | 1.762* | 0.743 | 0.925* | 0.396 | 0.786* | 0.385 | 0.778* | 0.364 |
| Exercise*Area | >4 times/DN | <once a week/DS | -4.573 * | 1.533 | -1.41 | 1.031 | -1.255 | 0.954 | -1.125 | 0.923 |
| | | 2 – 4 times/DS | -6.331 * | 1.627 | -1.981* | 0.913 | -1.805 * | 0.896 | -1.551 | 0.850 |
| Income*Breakfast | >R10000/daily | ≤R1000/not daily | -4.051 | 2.5 | -3.921* | 1.8 | -3.666 * | 1.731 | -3.808 * | 1.733 |
| | | R1001-R4500/not daily | 0.414 | 2.408 | -1.218 | 1.636 | -1.439 | 1.530 | -1.513 | 1.516 |
| | | R4501-R10000/not daily | 2.479 | 2.395 | -1.374 | 1.541 | -1.568 | 1.454 | -1.715 | 1.431 |
| TV*Breakfast | >3hrs/daily | ≤1hr/not daily | 6.310* | 2.213 | 2.573* | 1.259 | 2.051 | 1.192 | 1.976 | 1.186 |
| | | 1-3 hrs/not daily | 1.974 | 2.154 | 0.192 | 1.109 | 0.270 | 1.112 | 0.192 | 1.103 |
| SmokeVehicle*Age | no/ | yes/ | 0 | | 0.814* | 0.375 | 0.809* | 0.348 | 0.782* | 0.341 |
| Smoke preg*Area | no/DN | yes/DS | -5.118 * | 2.101 | -1.875 | 1.363 | -1.663 | 1.291 | 0 | |
| Work'nWear*Breakfast | no/not daily | yes/daily | 0 | | 2.349* | 1.076 | 2.095 | 1.070 | 2.165* | 1.090 |

DN – North Durban; DS – South Durban; preg – pregnant;

CC – Complete case

MVNI – Multiple imputed MVNI strategies 1 and 2

FCS1 -Multiple imputed FCS strategy 1

FCS2 -Multiple imputed FCS strategy 2

*Significant at the 0.05 level

A comparison of results of the complete case analysis(CC) with the other analyses showed that the standard errors of the estimated coefficients for the CC analysis are appreciably larger in all but the one predictor variable – 'smoke in vehicle'. There were also noticeable differences in the magnitude of the estimated coefficients for the CC analysis as compared to the other analyses. Contradictions were also present regarding the relationship with asthma severity for some of the predictors.

## 4.5 Diagnostics

In order to confirm that the imputed values are reasonable, each variable with missing data in excess of 8% was examined to identify variables with large differences between the measured and imputed. The variables considered included income, stove type, number of people and food availability (Figure 4.1). The Chi-square test was applied to assess whether significant differences exist between the distributions of the imputed data – both MVNI imputed and FCS imputed – and the measured data (Abayomi et al., 2008). No significant differences were found.

In analysis testing for significant differences between the distributions of the imputed data sets and the complete case data, no significant differences were found.

Another useful diagnostic that gives an indication of the stability of the estimates resulting from MI is the degrees of freedom (df) associated with the t-value in Rubin's rules and adapted from Schafer (1997) (Graham, 2012, Schafer and Olsen, 1998). The df associated with MI is not the same as the df found in other statistical concepts and rather is a 'measure' of the ratio of the within-imputation variance ($U$) to the between-imputation variance ($B$) such that

$$\mathrm{df} = (m-1)\left(1 + \frac{mU}{(m+1)B}\right)^2 \tag{4.1}$$

where $m$ = number of imputations. Thus the degrees of freedom are influenced by both the number of imputations and the relative sizes of $B$ and $U$. When $B$ dominates $U$ the degrees of freedom are close to the minimum value of $m$-1, but when $U$ dominates $B$ the degrees of

freedom approach infinity. If the computed value of df is very small (<10), it suggests that greater efficiency (more accurate estimates and narrower intervals) could be obtained by increasing the number of imputations, $m$. If df is large, however, it suggests that little will be gained from a larger $m$.

In this study, df ranged from 130.54 to 9073.51 for the NORM imputations and from 138.88 to 15135.431 for the FCS imputations which, being large compared to the number of imputed sets, is an indication that the estimates have stabilised and can be trusted.

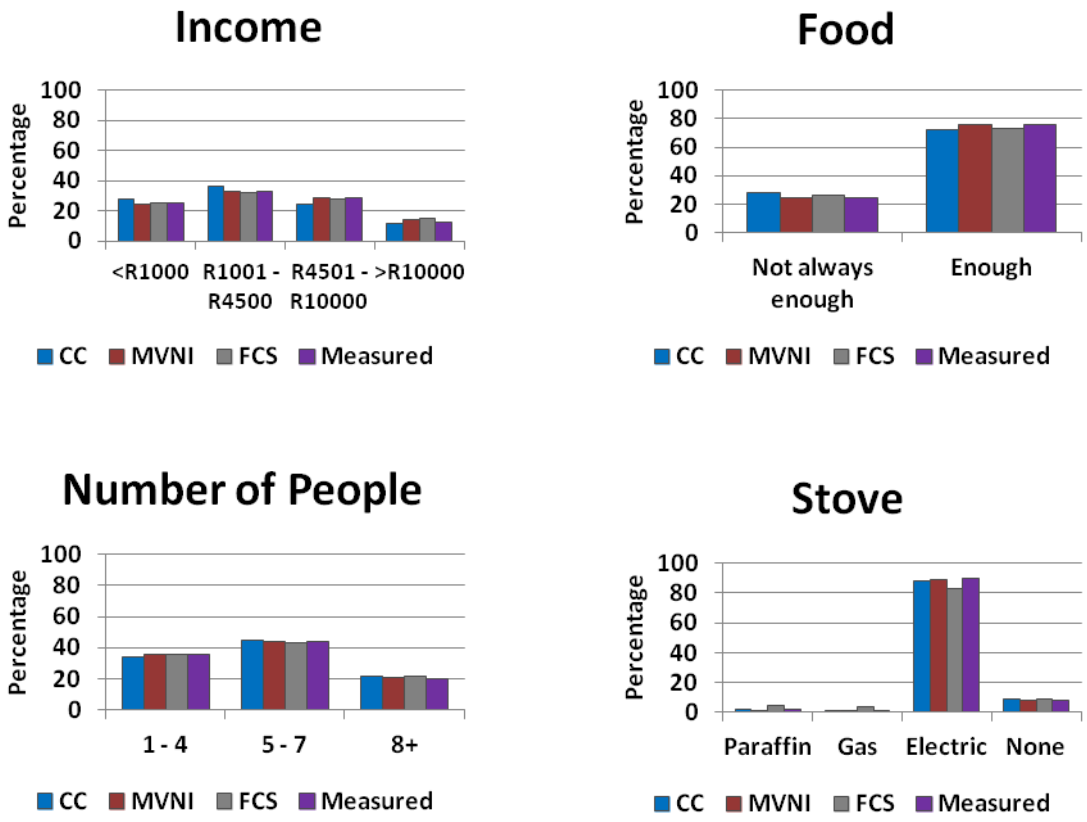



**Figure 4.1: Differences in measured (observed) and imputed data. A comparison of the distributions of the 4 variables with the most missing data for the complete case data (CC), MVNI imputed data, FCS imputed data and measured**

## 4.6 Discussion

In this investigation of identifying interactions in the presence of missing data, support was found for building the model using an EM generated set of data and then applying MI as a robust method to address this common shortcoming in epidemiological studies.

Epidemiological studies frequently suffer from missing data. Many researchers avoid this problem by dropping all cases with data missing on any variable and carrying out what is known as a complete case analysis. An advantage of this type of analysis is that it is computationally easy to apply and can be done with any reputable commercial software package. However, unless the data are MCAR, the values of the estimated coefficients produced with this analysis may be biased. Moreover, when the missingness is not only a function of the covariate(s) but also of the outcome variable, then the bias from a complete case analysis is heightened (Desai et al., 2011). Although complete case analysis and other *ad hoc* methods, like mean substitution and the missing-indicator method, are still widely used, researchers are becoming more aware of the perils of applying such methods and many are now employing MI methods to address the missingness in their data. While results from MI will be unbiased when data are MAR, it has been suggested that even when it is MNAR, adequately dealing with as much of the missingness mechanism as possible will usually produce sound results (Collins et al., 2001, Graham et al., 1997, Graham, 2012, Schafer and Olsen, 1998). This is achieved by including auxiliary variables – those variables related to the missingness but not necessarily included in the analysis, interactions and the outcome variable in the imputation model.

While much has been published on the application of MI to epidemiological studies, there is limited literature that deals with model building in the presence of missing data, and more specifically model building including interactions. The aim here was to demonstrate a simple and easily applied strategy to build interactions, which are not known up front, into a model while at the same time imputing the missing data.

The dilemma that was faced was a practical one. It is possible for the interactions to be added after imputation. This is termed passive imputation or ‘impute-then-transform’. However, it

has been shown that including interactions, as product terms, before imputation produces results superior to those achieved if the imputations are done first and the interactions are added at the analysis stage (Von Hippel, 2009). For the best results, the identified interactions should be included in the imputation model along with the predictor variables, the auxiliary variables and the outcome variable. However, how can the interactions be identified and the best model built, when the data are incomplete?

Two strategies for model building, S1 and S2, were explored – both utilising a single imputed data set generated from the ML parameter estimates produced from the EM algorithm for covariance matrices.

Imputation was carried out with both multivariate normal imputation (MVNI) and the more flexible fully conditioned specification (FCS). The same set of 17 predictor variables and 10 interactions for the best model were identified when applying MVNI with both strategies S1 and S2, as well as with the application of FCS and strategy S1. FCS with strategy S2 failed to include one of these predictors and an associated interaction in its best model. Since these dropped variables did not alter the interpretation of the results, it would seem that both strategies for model building are equally effective. The advantage of S1 over S2 is that it is easier and less time-consuming to execute and therefore probably the preferred choice.

In comparison to the model variables selected from the imputed data, fewer variables were selected for the model on the complete case data. This is most likely caused by the enormous reduction in cases and the subsequent loss of power.

A total of 5.3% missing items spread across 81.8% of variables, affecting 43.5% of cases was present in the dataset used for this analysis. Examination of the missingness revealed that it is possible that the missingness mechanism present in this data is a combination of MCAR, MAR and MNAR. Analysis of the relationships between both the missingness of the variables and the variables themselves confirmed that significant relationships exist between each of the variables and at least one other variable in the set; furthermore, the missingness of all but three of the variables is significantly related to at least one other variable in the set.

For reliable and unbiased results to be obtained from a complete case analysis, the data are required to be MCAR, which is clearly not the case here. Furthermore, although this means of dealing with missing data is acceptable when the lost cases amount to no more that 5%, this data set is reduced by over 40% which will inevitably have a negative effect on the outcome of the analysis.

On the other hand, MI, if applied correctly, is able to produce sound results when the data are MAR and it has been shown that even when the data are MNAR, the effects of this mechanism are often surprisingly minimal (Graham et al., 1997). In order to ensure that the imputation model was general enough to encompass the subsequent analysis, the outcome variable, interactions and variables related to either the incomplete variables, or their missingness, or both, were included in the imputation model. By including variables that are correlated with each incomplete variable but not its missingness, we expect that the additional information will cause a decrease in the standard errors and hence an increase in efficiency and statistical power (Collins et al., 2001). If there is an element of MNAR present in the data, the inclusion of these variables in the imputation model should lessen the bias and make the assumption of MAR more plausible.

It is unclear as to how many variables and interactions, given the sample size available, can be reliably assessed with MI applications. . It seems that this depends to some extent on the software being used. In some cases, convergence of large models is a problem in that it can make the imputation process unacceptably slow (White et al., 2011). Graham and Schafer (1999) , in a study using NORM to perform the imputations found that results were quite acceptable 'even with sample sizes as low as 50, even with as much as 50% missing from most variables, and even with relatively large and complex models'. In a study on the imputation of categorical data (Finch, 2010) it was found that, while problems exist when imputing using a variant of NORM designed to deal with categorical data when many variables are present, the same limitations are not problematic for NORM. In another study (Hardt and Görgen, 2008) on the inclusion of continuous auxiliary variables in the imputation model , the authors suggest the ratio of cases with complete data to variables should be at least 3:1. Given these guidelines, we found that convergence for both imputation methods was achieved quickly and reliably. Furthermore, even with the dummy coding of all the categorical variables and the

interactions, the ratio of complete cases to variables far exceeds 3:1. We are therefore confident that our results are reliable.

Diagnostic tests on the distributions of the imputed data showed that data imputed both with MVNI and FCS were not significantly different from either the measured data or the CC data. These results confirm findings that MI with MVNI incorporating sensible rounding should work in most situations (Schafer, 1997), even in the presence of binary and ordinal variables (Lee and Carlin, 2010).

The diagnostic measure, df, also indicated that the estimates obtained from both MI methods have stabilised and are therefore trustworthy.

Analysis of the two sets of imputed data yielded very similar results. This is consistent with findings from a study comparing the two imputation approaches (Lee and Carlin, 2010) where it was found that 'similar results can be expected from FCS and MVNI in a standard regression analysis involving variously scaled variables'. The magnitude of the standard errors and the magnitude and direction of the estimated coefficients were consistent across both these imputation types and for both model building strategies. While there were some inconsistencies in the significance of predictors, these did not affect the overall interpretation of the associations between asthma severity and the factors included on the models.

A comparison of results for the complete case analysis and the analyses of the imputed data showed that standard errors for the estimated coefficients from the analysis of the imputed data were, in all but one case, considerably smaller than those from the complete case analysis. These smaller standard errors resulted in greater accuracy of the estimated coefficients. This increased precision indicates the superior efficiency and statistical power obtained for the analysis of the imputed data. The inconsistencies in the signs of the estimates and the significance of the predictors could result from the non-random fashion in which cases are dropped for the complete case analysis which may distort the joint distribution among the variables. The resulting bias in point estimates could lead to misidentification of significant predictors (He, 2010). Another important factor that would negatively affect results of the complete case analysis is that the missingness mechanism present in the data is not confined

to being MCAR. While MI methods produce unbiased parameter estimates when the missingness is MAR, this is not the case with complete case analysis. This missingness mechanism factor could also have added to the large difference in magnitude of the standard errors for the complete case analysis as compared to the imputed data analysis that, some would argue, could not be explained on the basis of sample size alone.

These results are consistent with what one would expect given the significant reduction in cases for the complete case analysis and the missingness mechanism present in the data that would almost certainly result in a loss of power and the introduction of bias into estimates.

Given the rigid processes followed in the imputation of the data and subsequent analyses, it is suggested that the results from the imputed data can be considered reliable. On the other hand, the results from the complete case analysis should be treated with caution.

## 4.7 Conclusions

With the development of readily available and easily implemented software, MI methods for dealing with missing data are becoming more popular in epidemiological studies that have incomplete measured variables. A critical part of the imputation process is the inclusion of those variables that are correlated with missingness as well as the interactions to be used in the analysis process. While this can present a practical challenge if the interactions are not specified *a priori*, one possible approach has been illustrated that effectively identifies the best main effects and interactions for a model in the presence of missing data and at the same time, imputes the data items that are missing. Undoubtedly, further testing of these strategies on other data sets is needed. It is hoped that the ideas presented here can be further explored and developed so that, by addressing this practical dilemma, more medical researchers will be able to apply MI when data suffer from missingness.

# Chapter 5

# THE USE OF SUBSET CORRESPONDENCE ANALYSIS IN THE MANAGEMENT OF MISSING CATEGORICAL DATA

In Chapter 4, the management of missing data by means of MI was illustrated. This approach involved fitting the data to a model. Its application is restricted by complexities of models and distributional requirements. Many of the MI algorithms are more suited to dealing with missingness in continuous data. In this chapter, the application of subset correspondence analysis is investigated to address the issue of missingness in the analysis of categorical data. This method of analysis adopts a very different approach to the more traditional aforementioned method.

The theory of s-CA and how its output is interpreted was presented in Chapter 3. An application to the asthma data and the results are discussed in Section 5.3. For comparative purposes, some chi-square analyses were done to investigate the associations between asthma severity and individual variables (5.3.2).

## 5.1 Background

Missing categorical data is frequently encountered with survey data. Two methods commonly used to manage this form of missing data are complete case analysis – in which all records with incomplete data are excluded from analysis, and the 'indicator method' – in which an extra 'missing' category is added for each incomplete variable. These, and other ad hoc methods of dealing with missing data may, however, result in biased estimates and are thus not recommended (Greenland and Finkle, 1995, Little and Rubin, 1987) . A more acceptable tool that is becoming more popular for dealing with missing data, and often used in conjunction with some regression procedure to analyse multivariate data that suffer from missingness, is MI. This method of handling data is computationally complex and can be restrictive with its complexities of models and distributional requirements. An alternative approach is the

application of correspondence analysis (CA), and its variants, which is commonly used in the analysis of multivariate categorical data.

CA is primarily a graphical technique used to explore the relationships between variables. Unlike the more classical regression-based methods for studying inter-variable relationships which hypothesise a model and fit the data to a model, the extended family of methods under CA do not hypothesise a model. Instead, the data are decomposed in order to study their 'structure' (Greenacre, 1984). Points (rows and columns of a data matrix), represented as clouds in multi-dimensional space, are optimally displayed in a lower dimensional subspace that is easier to interpret due to the lower dimensionality. The development of s-CA has made it possible to analyse a subset of the original data. This can be applied to data that suffer from missingness. The non-response for each variable is categorised separately and the subset of observed categories is analysed. This method offers a way of dealing with missing categorical data while, at the same time, retaining all records, complete and incomplete.

## 5.2 Preparation of the data

The same set of 22 variables that was used in the application of MI (Chapter 4) is used for this analysis. Details of these variables and their categories can be found in Hendry et al. (2014a). Because CA requires that data be categorical, the continuous variable 'age' is categorised into a 4-level variable – 8-9 years; 10 years; 11 years; and 12+ years. These data are represented in the form of a contingency table (Table 5.1) with four rows –representing the four asthma categories – and 71 columns – representing the categories of the 21 remaining variables plus a separate missing category for each variable that suffered from non-response.

All missing categories are excluded from the subset for analysis. Also excluded is the category BW? - of the birth weight variable. This category is a response option for respondents who did not know the birth weight of the child, and it is considered to play a similar role to BW* (non-response to the birth weight question).

**Table 5.1: Contingency table (split up) showing frequencies of variables across asthma severity categories**

| | A1 | A2 | A3 | A4 | MAL | FEM | BW1 | BW2 | BW? | BW* | NNY | NNN | NN* | FrY | FrN | Fr* | SPY | SPN | SP* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ASMS** | 5 | 13 | 8 | 1 | 20 | 7 | 6 | 18 | 3 | 0 | 9 | 18 | 0 | 10 | 15 | 2 | 2 | 24 | 1 |
| **ASMP** | 7 | 22 | 14 | 4 | 17 | 30 | 10 | 32 | 5 | 0 | 9 | 36 | 2 | 17 | 26 | 4 | 5 | 40 | 2 |
| **ASMI** | 4 | 39 | 26 | 7 | 36 | 40 | 7 | 62 | 7 | 0 | 7 | 68 | 1 | 39 | 33 | 4 | 7 | 67 | 2 |
| **ASN** | 9 | 122 | 87 | 14 | 90 | 142 | 33 | 168 | 27 | 4 | 25 | 196 | 11 | 99 | 118 | 15 | 21 | 197 | 14 |

| | SY | SN | S* | SVY | SVN | SV* | E1 | E2 | E3 | E* | T1 | T2 | T3 | T* | N1 | N2 | N3 | N* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ASMS** | 15 | 12 | 0 | 8 | 17 | 2 | 8 | 7 | 9 | 3 | 12 | 10 | 3 | 2 | 13 | 9 | 3 | 2 |
| **ASMP** | 25 | 22 | 0 | 13 | 30 | 4 | 14 | 17 | 12 | 4 | 8 | 26 | 8 | 5 | 19 | 17 | 6 | 5 |
| **ASMI** | 35 | 41 | 0 | 18 | 53 | 5 | 21 | 27 | 24 | 4 | 18 | 38 | 16 | 4 | 20 | 34 | 13 | 9 |
| **ASN** | 112 | 119 | 1 | 55 | 159 | 18 | 70 | 84 | 65 | 13 | 48 | 119 | 51 | 14 | 72 | 93 | 48 | 19 |

| | I1 | I2 | I3 | I4 | I* | Fne | Fe | F* | O | C | U | PW* | WWY | WWN | WW* | PY | PN | P* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ASMS** | 5 | 6 | 10 | 2 | 4 | 16 | 7 | 4 | 2 | 5 | 18 | 2 | 5 | 22 | 0 | 4 | 23 | 0 |
| **ASMP** | 9 | 13 | 13 | 4 | 8 | 29 | 10 | 8 | 5 | 3 | 35 | 4 | 8 | 37 | 2 | 9 | 36 | 2 |
| **ASMI** | 24 | 17 | 11 | 8 | 16 | 50 | 19 | 7 | 10 | 8 | 53 | 5 | 6 | 69 | 1 | 20 | 56 | 0 |
| **ASN** | 41 | 66 | 54 | 25 | 46 | 170 | 49 | 13 | 37 | 19 | 161 | 15 | 17 | 204 | 11 | 81 | 149 | 2 |

| | DS | DN | Bnd | Bd | B* | VY | VN | V* | WY | WN | W* | p | g | e | n | St* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ASMS** | 18 | 9 | 17 | 8 | 2 | 9 | 15 | 3 | 9 | 15 | 3 | 1 | 0 | 22 | 2 | 2 |
| **ASMP** | 27 | 20 | 31 | 11 | 5 | 19 | 24 | 4 | 15 | 27 | 5 | 2 | 0 | 35 | 4 | 6 |
| **ASMI** | 39 | 37 | 45 | 26 | 5 | 41 | 31 | 4 | 39 | 33 | 4 | 0 | 1 | 60 | 5 | 10 |
| **ASN** | 103 | 129 | 143 | 76 | 13 | 116 | 99 | 17 | 97 | 119 | 16 | 3 | 2 | 191 | 16 | 20 |

## 5.3 Results

### 5.3.1 Subset correspondence analysis

With the application of this dataset, the objective was to identify relationships between the environmental, socio-economic, genetic and behavioural variables and to investigate possible relationships between these variables and asthma severity. CA was initially applied to the full data set in which missing categories were present. The total inertia amounted to 0.0207.

It was found that a number of the non-response categories contributed highly to the orientation of axis 2. This resulted in an elongation of the scale along this axis which, in turn, resulted in a clumping together of variables near the origin. This made it very difficult to distinguish between the points and interpret the maps, and masked more relevant relationships in the data. Furthermore, given the large number of variables in the data set, the inclusion of the non-response categories exacerbated the situation of an already crowded display. To address these phenomena, s-CA was applied to the subset of observed data, thus excluding the non-response categories from the analysis.

The missing data accounted for 21.7% of the variability in the full data set (Hendry et al., 2014a). The remaining 78.3%, which is the variability of the measured data, was further decomposed in the analysis of the subset of observed data. The first two axes of the analysis of this subset accounted for 88.92% of the total inertia. By examining the decomposition of inertia along these axes (Table 5.2), interpretation of the principal axes is possible.

For axis 1, the variables that made the most contribution to the orientation of this axis are A1 (age 8 – 9 years) and NNY (having received some form of special neo-natal care). Both physiological variables have been separated out from the other variables and are situated on the negative side of the axis. Other variables that contributed to this axis and are associated with the aforementioned variables are WWY (exposure to secondary smoke and chemicals), male, N1 (up to 4 people in the home), I3 (income of R4501 – R10000), T1 (<1hr TV a day), DS (from South Durban), p (those who use a paraffin stove) and BW1 (<2.5kg at birth). Opposing these, on the positive side, are PY (having had a pet), DN (from North Durban) and

**Table 5.2: Decomposition of inertia for the first 2 principal axes**

| Name | Mass | QLT | INR | k= 1 | COR | CTR | k= 2 | COR | CTR |
|---|---|---|---|---|---|---|---|---|---|
| A1 | 3 | 944 | 3 | -717 | 943 | 149 | -28 | 1 | 1 |
| A2 | 24 | 719 | 0 | 34 | 712 | 3 | 3 | 7 | 0 |
| A3 | 17 | 750 | 0 | 73 | 725 | 8 | -13 | 25 | 1 |
| A4 | 3 | 154 | 0 | 57 | 59 | 1 | 72 | 95 | 5 |
| MAL | 20 | 881 | 2 | -149 | 455 | 42 | 144 | 426 | 116 |
| FEM | 27 | 881 | 1 | 111 | 455 | 31 | -107 | 426 | 86 |
| BW1 | 7 | 997 | 1 | -220 | 673 | 31 | -153 | 324 | 45 |
| BW2 | 35 | 907 | 0 | 37 | 350 | 4 | 46 | 557 | 21 |
| NNY | 6 | 974 | 0 | -475 | 974 | 131 | -10 | 0 | 0 |
| NNN | 40 | 958 | 68 | 65 | 882 | 16 | 19 | 76 | 4 |
| FrY | 21 | 974 | 18 | 68 | 395 | 9 | 83 | 579 | 39 |
| FrN | 24 | 944 | 19 | -46 | 370 | 5 | -58 | 574 | 22 |
| SPY | 4 | 257 | 3 | 13 | 28 | 0 | -37 | 229 | 2 |
| SPN | 41 | 986 | 18 | -7 | 152 | 0 | 16 | 834 | 3 |
| SY | 23 | 998 | 0 | -51 | 862 | 6 | -20 | 136 | 3 |
| SN | 24 | 991 | 32 | 47 | 816 | 5 | 22 | 175 | 3 |
| SVY | 12 | 985 | 8 | -76 | 975 | 6 | -8 | 10 | 0 |
| SVN | 32 | 989 | 4 | 30 | 870 | 3 | 11 | 119 | 1 |
| E1 | 14 | 737 | 17 | -2 | 2 | 0 | -29 | 735 | 3 |
| E2 | 17 | 824 | 5 | 60 | 666 | 6 | -29 | 158 | 4 |
| E3 | 14 | 934 | 14 | -15 | 41 | 0 | 71 | 893 | 19 |
| T1 | 11 | 751 | 18 | -187 | 445 | 35 | 155 | 306 | 71 |
| T2 | 24 | 634 | 11 | 45 | 326 | 5 | -44 | 308 | 13 |
| T3 | 10 | 999 | 92 | 147 | 985 | 19 | -18 | 14 | 1 |
| N1 | 15 | 986 | 16 | -169 | 890 | 41 | -56 | 96 | 13 |
| N2 | 19 | 940 | 2 | 61 | 634 | 7 | 42 | 306 | 9 |
| N3 | 9 | 820 | 14 | 161 | 779 | 21 | -37 | 41 | 3 |
| I1 | 10 | 722 | 3 | 38 | 21 | 1 | 221 | 701 | 132 |
| I2 | 13 | 951 | 13 | 34 | 118 | 1 | -91 | 833 | 29 |
| I3 | 11 | 836 | 1 | -189 | 612 | 36 | -114 | 224 | 40 |
| I4 | 5 | 970 | 134 | 102 | 970 | 5 | 2 | 0 | 0 |
| Fne | 33 | 763 | 29 | 59 | 646 | 11 | -25 | 117 | 6 |
| Fe | 11 | 989 | 9 | -32 | 150 | 1 | 75 | 839 | 17 |
| O | 7 | 914 | 22 | 169 | 832 | 18 | -53 | 82 | 5 |
| C | 4 | 814 | 88 | -187 | 363 | 14 | 208 | 451 | 52 |
| U | 33 | 260 | 32 | -3 | 9 | 0 | -13 | 251 | 2 |
| WWY | 4 | 928 | 0 | -406 | 911 | 69 | -55 | 17 | 4 |
| WWN | 41 | 944 | 42 | 35 | 663 | 5 | 23 | 281 | 6 |
| PY | 14 | 816 | 0 | 200 | 739 | 53 | -65 | 77 | 16 |
| PN | 33 | 887 | 26 | -80 | 702 | 20 | 41 | 185 | 15 |
| DS | 23 | 932 | 26 | -126 | 845 | 34 | 40 | 87 | 10 |
| DN | 24 | 932 | 16 | 120 | 845 | 33 | -39 | 87 | 10 |
| Bnd | 29 | 952 | 1 | -19 | 377 | 1 | -23 | 575 | 4 |
| Bd | 15 | 766 | 2 | 71 | 468 | 7 | 56 | 298 | 13 |
| VY | 23 | 994 | 28 | 111 | 912 | 26 | 33 | 82 | 7 |
| VN | 21 | 999 | 17 | -95 | 945 | 18 | -23 | 54 | 3 |
| WY | 20 | 995 | 22 | 99 | 492 | 18 | 100 | 503 | 55 |
| WN | 24 | 985 | 6 | -50 | 358 | 6 | -66 | 627 | 29 |

**Table 5.2: Decomposition of inertia for the first 2 principal axes (continued)**

| Name | Mass | QLT | INR | k= 1 | COR | CTR | k= 2 | COR | CTR |
|---|---|---|---|---|---|---|---|---|---|
| p | 1 | 983 | 15 | -689 | 671 | 33 | -470 | 312 | 45 |
| g | 0 | 999 | 96 | 453 | 707 | 7 | 291 | 292 | 9 |
| e | 38 | 174 | 1 | 14 | 173 | 1 | 1 | 1 | 0 |
| n | 3 | 788 | 2 | -55 | 451 | 1 | -47 | 337 | 2 |
| ASMS | 71 | 951 | 85 | -312 | 909 | 638 | 67 | 42 | 88 |
| ASMP | 123 | 777 | 148 | -131 | 563 | 195 | -80 | 214 | 220 |
| ASMI | 199 | 887 | 197 | 45 | 142 | 38 | 103 | 745 | 585 |
| ASN | 607 | 865 | 570 | 48 | 676 | 130 | -25 | 189 | 108 |

Mass (Mass) and inertia (INR) of each variable; the quality (QLT) of the variable's representation in the subspace of the first 2 axes; Co-ordinates (k = ...); contributions of axes to the inertia of the variables (COR); and contributions of variables to the inertia of the axes (CTR)*.
*For details of the formulae for calculations see Greenacre (1984), p 91.

female. Moderate to severe (ASMS) and mild persistent (ASMP) asthma are associated with the groupings on the negative side and 'no asthma' (ASN) with the group on the positive side.

Many variables did not play a major role in the orientation of the axis but are correlated with it, as evidenced by the large COR values. In particular, the smoke exposure variables, both in the home (SY) and in vehicles (SVY), are highly correlated with this axis and are situated on the negative side indicating an association with the more severe levels of asthma.

It is evident that subjects are separated on this axis on the basis of both physiographic factors and smoke exposure. These were the biggest contrasts in the data and accounted for 66.52% of the total inertia.

The orientation of axis 2 was defined mainly by the variables I1 (income of <R1000), male and female, T1 (less than 1 hour TV a day) and WY (being attacked with weapons). There was a separation on this axis of those subjects who: are from the lowest income group (I1); are male; experience fear in the neighbourhood (FrY); have been attacked with weapons (WY) and watch TV for less than an hour a day (T1); from those subjects who: are female; have not been

attacked with weapons (WN); are from the R1001 – R4500 income group (I2) and do not experience fear in the neighbourhood (FrN). The mild intermittent asthma variable (ASMI) is correlated with the former grouping. Axis 2 can be thought of as distinguishing between subjects on the basis of their socio-economic status (SES) and accounted for 22.4% of the total inertia.

When interpreting the graphical display, those variables that are not well represented in the subspace are situated near the origin and do not add to the interpretation of the display. By examining the angles that the points make with each other and with the principal axes, trends and relationships present in the data could be identified and interpreted.

In the plane of the first and second axis (Figure 5.1), which accounted for 88.9% of the variation in the data, the physiological/smoke exposure axis was plotted against the socio-economic axis. Variables indicative of low socio-economic status are situated above the horizontal axis and the higher socio-economic variables below. In the same way, the vertical axis separates the smoke exposure variables as well as those representing low birth weight (BW1); having had neo-natal care (NNY); male and low age (A1) from their 'opposites'. The asthma variables are well represented in this subspace. The more severe asthma variables (ASMS and ASMP) are split from the other categories (ASMI and ASN) by the vertical axis indicating an association of worse asthma with those variables situated to the left of the axis. Mild intermittent asthma (ASMI) is removed from the other three asthma variables and tends in the direction of lower socio-economic status. Further distinctions between the levels of asthma severity are evidenced by their locations – each in a different quadrant.

The strongest associations with moderate to severe asthma (ASMS) were shown by males, having had neo-natal care (NNY), smoke exposure in vehicles (SVY), 8-9 year olds (A1) and coming from South Durban (DS); mild persistent asthma was associated most with a birth weight of less than 2.5kg (BW1), using a paraffin stove (p) or not having a stove (n), smoke exposure in the home (SY), exposure to secondary smoke and chemicals (WWY), living in a home with up to 4 people (N1) and a monthly income of R4501 – R10000 (I3); and associations with mild intermittent asthma were shown by the lowest income group (I1), a birth weight of more than 2.5kg (BW2), being attacked by weapons (WY), experiencing fear in the neighbourhood (FrY) and doing exercise more than 4 times a week (E3).

An interesting result is the distinction between the different forms of smoke exposure and their associations with asthma severity. A close association is evident between smoke exposure in the home (SY) and mild persistent asthma (ASMP). Smoke exposure in a vehicle (SVY) shows a stronger association with moderate to severe asthma (ASMS) than with mild persistent asthma (ASMP), as indicated by the angles that the point vectors make with the asthma variables. Exposure to severe levels of air pollution, as experienced in the South Durban region (DS), shows a strong association with moderate to severe asthma (ASMS). Smoking while pregnant (SPY) is not well represented in this subspace and is therefore not included in this discussion.

Another interesting phenomenon is the positioning of the stove variables, paraffin (p) and gas (g), at opposite corners of the display. The association of gas stove (g) with mild intermittent asthma (ASMI) contrasts that of paraffin stove (p) with mild persistent asthma (ASMP).

With regard to the number of people in the home and its association with asthma severity, results show that N1 (1-4 people) tends in the direction of mild persistent asthma (ASMP), N2 (5-7 people) tends towards mild intermittent asthma (ASMI) and N3 (8+ people) tends towards no asthma (ASN). Thus there is an inverse relationship between asthma severity and the number of people there are in the home.

It can be seen that the inertia associated with this subspace amounts to 0.0144 (0.0108 + 0.0036) in total. This relatively low value indicates that there is not a lot of variability in the data and explains the bunching up of the variables in the display (Greenacre, 1992).

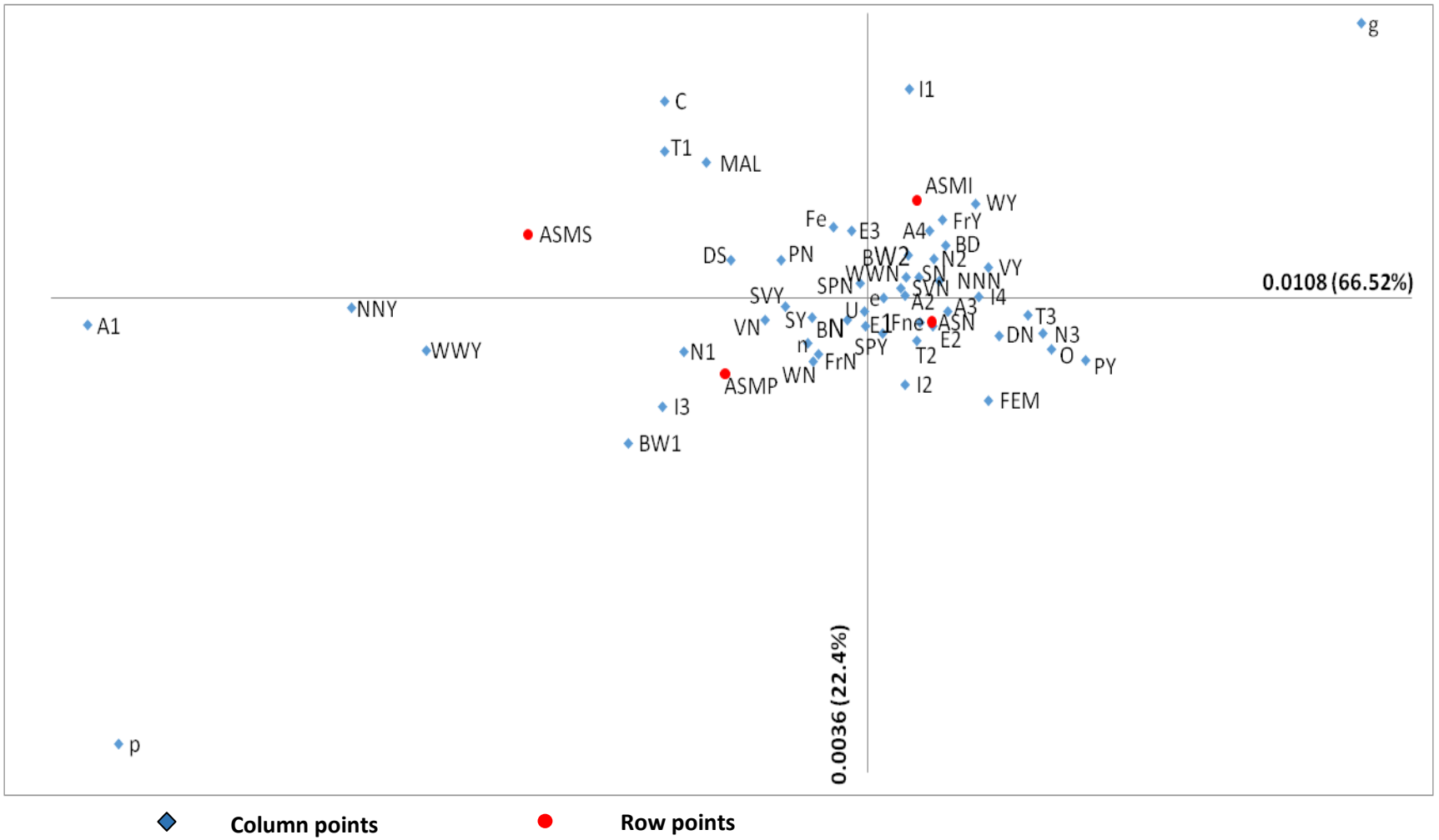



**Figure 5.1: Subset CA map of a contingency table with the row and the column points projected onto the plane of the first and second principal axes. Values on the axes indicate principal inertias and their respective percentages of total inertia.**

### 5.3.2 Chi-square analysis

As a comparative method of association analysis of contingency tables, Pearson's chi-square test was applied to individual cross-tabulations of asthma severity with each of the generic, socio-economic, behavioural and environmental variables. By examining the contributions of individual cells to the chi-square value, we were able to identify specific relationships between the two variables in the table. In addition, Cramer's V statistic gave us an indication of the relative strength of the associations found.

Results from Pearson's chi-square (Table 5.3) showed that there was agreement, at the 5% level of significance, that gender, neo-natal care and ever having pets are significantly related to asthma severity. Specifically, significantly more than expected of the subjects who were male or who had had specialist care at birth and significantly fewer than expected of those who ever had pets suffered from moderate to severe asthma. Relaxing the level of significance to 10%, associations were found to exist between asthma severity and age, area and exposure to secondary chemicals and dust. More specifically, more than expected of the youngest age group as well as those who were exposed to secondary chemicals and smoke suffered from moderate to severe or mild persistent asthma; while more than expected of those from South Durban had moderate to severe asthma.

**Table 5.3: Results of Pearson's chi-square and Cramer's V tests for the 10 variables that exhibit the strongest relationship with asthma**

| Variable (categories) | Chi-square p-value | Worse asthma associated with... | Cramer's V |
|---|---|---|---|
| Gender (MAL/FEM) | 0.003 | MAL | 0.190 |
| Neo-natal care (NNY/NNN)) | 0.005 | NNY | 0.186 |
| Pets (PY/PN) | 0.036 | PN | 0.150 |
| Work and wear (WWY/WWN) | 0.070 | WWY | 0.138 |
| Area (DS/DN) | 0.077 | DS | 0.134 |
| Age (A1/A2/A3/A4) | 0.087 | A1 | 0.120 |
| TV (T1/T2/T3) | 0.137 | TV1 | 0.117 |
| Birth weight (BW1/BW2) | 0.182 | BW1 | 0.120 |
| Income (I1/I2/I3/I4) | 0.216 | I3 | 0.114 |
| Weapons (WY/WN) | 0.217 | WN | 0.112 |

Cramer's V statistic (Table 5.3) indicates that the three strongest associations are exhibited between asthma severity and gender, neo-natal care and pets, respectively. While the values of this statistic signify only a low association for each of the variables shown, they are large enough to suggest that a relationship between asthma severity and each of these variables does exist.

## 5.4 Discussion

In this application of s-CA to a dataset with a substantial amount of missing data, the use of this technique was shown to provide a meaningful approach to exploring the relationships between categorical variables that suffer from missingness. This approach provides several advantages when compared to other methods of addressing such shortcomings of data sets. The advantages are that the method is not constrained by either model assumptions or distributional requirements; it is computationally simple; and it is able to handle large numbers of categorical variables. All the standard analysis was performed using SPSS (version 17), and a macro program was written to perform the s-CA.

Applying CA to the full data set resulted in an elongation of the scale on axis 2, which exacerbated an already crowded display, thus making it difficult to identify points and interpret relationships between them. In addition, it is the relationships between the measured variables and level of asthma severity that are of interest in this study. Due to the useful property of s-CA, whereby the full data matrix can be partitioned into smaller mutually exclusive sub-matrices, with the respective decomposition of the total inertia, CA was applied to the sub-matrix of observed variable categories only, which allowed for a clearer display of the points and enabled the exploration of the relationships between the relevant variables.

The application of this novel explorative statistical technique has enabled us to examine a large number of environmental, behavioural, genetic and socio-economic variables to uncover relationships between these variables and, at the same time, retain all records. Furthermore, associations between these variables and asthma have been found that generally confirm established theories regarding factors that exacerbate asthma. We have further been able to

distinguish between different levels of asthma severity and the factors that are associated with them.

There is agreement that asthma is associated with younger children (Asher et al., 2006) ; a birth weight of less than 2.5kgs and having had neo-natal care (Mai et al., 2003); exposure to low concentrations of compounds and pollutants as a result of living in the same house with someone who works in a chemical/dust environment and wears their work clothes home (Becher et al., 1996, Venables and Chan-Yeung, 1997); male children (Almqvist et al., 2007, Bonner, 1984) and smoke exposure both in vehicles (Sendzik et al., 2009) , in the home (Charoenca et al., 2013, Ehrlich et al., 1992) and in the form of air pollution (Neidell, 2004, Peden, 2005). These variables are shown to be associated with the higher levels of asthma severity in this application. The counter-intuitive association found between having pets and suffering from mild/no asthma does not contradict international findings that yield conflicting results on the association between pets and asthma.

Other studies that have led to results that confirm documented theories for factors that influence asthma severity include: that the risk from exposure to smoke in a car exceeds the risk from smoke in the home (Sly et al., 2007); that there is an association between asthma and indicators of low SES, viz. experiencing fear in the neighbourhood (Subramanian and Kennedy, 2009); neighbourhood stressors in the form of the use of weapons (Jeffrey et al., 2006, Wright et al., 2004); and low income homes (Cesaroni et al., 2003, Poyser et al., 2002); and that asthma occurrence is inversely related to the size of the family (Matricardi et al., 1998). This last, perhaps unexpected relationship, could result from the possibility that common infections acquired early in infancy because of unhygienic contacts with older siblings, could better 'protect' from atopic diseases like asthma (Strachan, 1989).

Relative weights and inter-point distances are retained from the analysis of the full data set and are not recalculated for the analysis of the subset. This allows for the decomposition of the inertia into parts representing mutually exclusive and exhaustive subsets. CA of the full data set resulted in a total inertia of 0.0207. This is a measure of the dispersion of the points in the full m-dimensional space. The analysis of the subset of observed categories yielded a total inertia of 0.0162 and total inertia from the analysis of the non-response categories is 0.0045. Due to the fact that the two subsets are mutually exclusive and exhaustive, the sums of their

total inertias equal the total inertia of the whole data set. Furthermore, the observed categories account for nearly four times as much of the inertia (0.0162 / 0.0207 = 78.3%) in the data as do the non-response categories (0.0045 / 0.0207 = 21.7%). While we have been able to identify many interesting relationships in the data, we can see from the correspondence map that the dispersion of the points is not extensive. This is borne out by the value of the total inertia (a relatively low 0.0162), which is a measure of how much the measured profiles are spread around the origin. The low variability present in the data is further confirmed by the small number of variables that show a significant relationship with asthma as demonstrated with the chi-square association analysis.

While it is important to note that, with s-CA, relationships found to exist between variables/categories cannot be assumed to be statistically significant, comparative tests of association were carried out on cross-tabulations of asthma severity with the other variables. Relationships between asthma severity and a number of the variables included in the study were identified. Despite the fact that the associations were not necessarily strong, they do corroborate the associations found with s-CA. The fact that only a few variables were found to be significantly associated with asthma severity is consistent with our finding in s-CA that the dispersion of points was not large, as seen both in the graphical display and in the low inertia value.

It has thus been shown that s-CA, as presented here, has a two-fold purpose: firstly, as an exploratory tool to seek inter-relationships between variable categories and to identify those variable categories that are associated with different levels of childhood asthma so that they can be taken further and used in more rigid analysis; and secondly, to manage the missing data and the problem of crowding created by it. Furthermore, where large numbers of variables/categories are involved, relationships between variables/categories are not generally easy to summarise. So this could be taken a step further and subsequent division of the data into numerous smaller, sensibly selected, mutually exclusive and exhaustive subsets is suggested. In these situations, it is proposed that s-CA is an ideal choice of method and produces easily interpreted graphical output to provide a general view of the associations between the many variables.

## 5.5 Conclusions

Despite the presence of missing data, s-CA is able to explore the data as a whole and represent the variables graphically, thus implying relationships between variables. By identifying those variables important to the determination of the principal axes, the identification of a selection of the variables to take forward for further analysis is possible. It is believed that this exploratory method is easier to apply than the existing MI methods in which many complexities need to be considered. While MI allows one to carry out statistical analysis on data that encounter missingness, the sophistications in the assumptions about the model, missingness mechanisms and computational algorithms are restrictive and make it more difficult to use. The s-CA approach offers an alternative paradigm to dealing with the analysis of categorical data that suffer from missingness.

# Chapter 6

# THE EFFECT OF THE MECHANISM AND AMOUNT OF MISSINGNESS ON SUBSET CORRESPONDENCE ANALYSIS

The use of s-CA to manage missing categorical data was presented in Chapter 5. Its application is a relatively new technique in the handling of missing categorical data. While many studies have examined the effects of missingness mechanisms and the amount of missingness on MI (Hardt and Görgen, 2008, Marshall et al., 2010, Peyre et al., 2011, Shrive et al., 2006), it is not known what effect these factors have on s-CA. In this chapter a simulation study is presented that tests the effects of Little and Rubin's missingness mechanisms as well as missingness of up to 50% on the analysis of data using s-CA. An outline of the different scenarios tested is given in Section 6.2 while Section 6.3 introduces the outcomes used to measure the effects of the mechanisms on s-CA. Results from the analysis of a full set of data, used as a benchmark analysis, are presented in Section 6.4 followed by the results from the simulation study (Section 6.5).

## 6.1 Selected variables

From the original 22 variables included in this study, a purposeful selection of six variables was made. These included 'age' (categorized into 4 levels from 9 years to 12+ years); 'gender' (M/F); 'neo-natal' (whether or not special neo-natal care was received at birth); 'smokers' (the presence of smokers in the home) and 'area' (North or South Durban) and the 'asthma severity' variable (none/mild intermittent; mild persistent; moderate to severe). These variables were chosen because, in previous analyses (Chapter 4 and Chapter 5), they were shown to have some association with asthma severity.

For the purpose of this investigation, the asthma severity variable was condensed into three categories, as was used in the analysis with MI. Thus 'mild persistent' and 'no asthma' were combined. This was done to facilitate interpretation of results since the set of data thus formed is fully represented in a 2-dimensional subspace.

All child records with complete data on all six variables were included in this study. This enabled a benchmark analysis to be made in which no missing data are present. A summary of this data, involving 368 cases, can be found in Table 6.1.

**Table 6.1: Categories, code names and frequencies for all variables**

| Variables | Categories | Code names | Count (N = 368) |
|---|---|---|---|
| Age | 8 - 9 years | A1 | 24 |
| | 10 years | A2 | 186 |
| | 11 years | A3 | 134 |
| | 12+ years | A4 | 24 |
| Gender | Male | MAL | 149 |
| | Female | FEM | 219 |
| Neo-natal | Yes | NNY | 50 |
| | No | NNN | 318 |
| Smokers | Yes | SY | 180 |
| | No | SN | 188 |
| Area | South Durban | DS | 177 |
| | North Durban | DN | 191 |
| Asthma severity | None/ Mild intermittent | ASNI | 296 |
| | Mild persistent | ASMP | 45 |
| | Moderate/severe | ASMS | 27 |

## 6.2 Missing data mechanisms

To explore the effect of missingness mechanisms (MM) and amount of missingness present (M%), 18 scenarios were considered, with each scenario simulated 10 times. Three MM's were imposed – MCAR, MAR and MNAR – and missingness was generated at rates of 5%, 10%, 20%, 25%, 30% and 50% for each mechanism. Two variables – 'neo-natal' and 'smokers' – were selected to experience missingness and the same amount of data were deleted from each of these variables for each scenario.

For the six MCAR scenarios, data were deleted randomly across all categories for each of the variables.

To simulate the MAR mechanism, missingness was imposed on the 'neo-natal' and 'smoking' variables according to their association with 'area' and 'gender' respectively. Data were randomly deleted from the 'neo-natal' variable such that 30% came from North Durban and 70% from South Durban. This ratio was selected to mimic the missingness of 'neo-natal' with respect to 'area' in the full data set. Random deletion on the 'smoker' variable was in the ratio 30:70 for M:F. Since there was only one missing data item for this variable, there was nothing to guide this deletion, so this was a subjective choice. These deletions were completed for each of the six amounts of missingness.

The MNAR mechanism was simulated so that the missing data depended on the actual value of the data item. Deletion from the 'neo-natal' variable was carried out such that 10% of required deletions were from the variable category NNY and 90% from variable category NNN. In a similar manner, deletions from the 'smoker' variable involved randomly deleting 90% of required deletions from SY and 10% from SN. These deletion ratios were considered to be sensible, given the setting. Again this was repeated for the six amounts of missingness.

## 6.3 Outcomes of interest

s-CA was applied to each of the simulated data sets and several outcomes were examined to identify effects of the MM and M% on this method. These included:

COR - relative contributions that the axis makes to the inertia (variance) of the points

CTR - absolute contributions that the points make to the inertia of the axis

TOTINR - a measure of the degree of variation in the measured data

TI%FULL- the proportion that TOTINR is of the total inertia from an analysis which includes both the measured and the missing data, coded as separate 'missing' categories

Repeated measures ANOVA was applied to the above outcomes to test for significant differences across missingness mechanisms and amount of missingness.

## 6.4 Full analysis

For the purpose of comparison, s-CA was applied to the 368 data set with all variables fully measured. The data were in the form of a contingency table with the three asthma categories as rows and the five selected variables (12 variable categories) as columns.

Results (Figure 6.1 and Table 6.2) showed that total inertia across the full subspace of two dimensions is 0.0271, thus indicating that there is limited variability in the data.

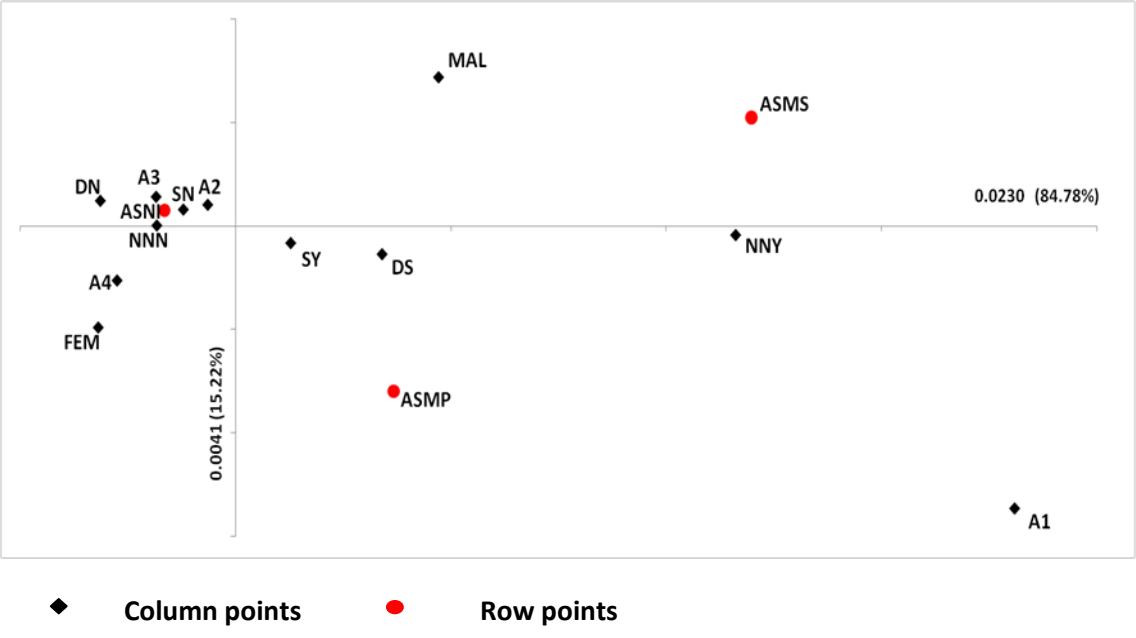



**Figure 6.1: Subset correspondence analysis map of the completely measured 368 data set. Values on the axes represent principal inertias and their respective percentages of total inertia. Labels as specified in Table 6.1.**

CTR values, a measure of the absolute contributions of the points to the inertia of the dimension, indicated that variable categories important to the orientation of axis 1 are A1, MAL, NNY and to a lesser extent FEM, DS and DN. This axis separated the lowest asthma

severity category (ASNI) on the left from the higher asthma severity categories (ASMP and ASMS) on the right. Associated with the latter categories are A1, MAL, NNY and DS. Variables that played an important part in the orientation of axis 2 are A1, MAL and FEM. This axis separated out ASMS from the other asthma categories, thus enabling a distinction between ASMP and ASMS. Associated with ASMP are FEM and A1.

Variables that do not exhibit much variance are situated near the origin. They do not play an important role in the orientation of the axes. These included A2, A3, A4, SY, SN, NNN and ASNI. Associated with the lowest asthma severity classification were A2, A3, SN, DN and NNN.

COR values indicated that axis 1 is more important in terms of contributions to inertia for all variable categories, except ASMP.

**Table 6.2: Decomposition of inertia for the two principal axes**

| Name | k = 1 | COR | CTR | k = 2 | COR | CTR |
|---|---|---|---|---|---|---|
| A1 | 723 | 875 | 297 | -273 | 125 | 236 |
| A2 | -26 | 606 | 3 | 21 | 394 | 10 |
| A3 | -74 | 864 | 17 | 29 | 136 | 15 |
| A4 | -110 | 819 | 7 | -52 | 181 | 8 |
| MAL | 188 | 631 | 125 | 144 | 369 | 407 |
| FEM | -128 | 631 | 85 | -98 | 369 | 277 |
| NNY | 464 | 1000 | 255 | -8 | 0 | 0 |
| NNN | -73 | 1000 | 40 | 1 | 0 | 0 |
| SY | 51 | 910 | 11 | -16 | 90 | 6 |
| SN | -49 | 910 | 11 | 16 | 90 | 6 |
| DS | 136 | 961 | 78 | -27 | 39 | 18 |
| DN | -126 | 961 | 72 | 25 | 39 | 16 |
| ASNI | -66 | 952 | 153 | 15 | 48 | 43 |
| ASMP | 147 | 458 | 116 | -160 | 542 | 762 |
| ASMS | 479 | 954 | 732 | 105 | 46 | 195 |

K=… co-ordinates

COR relative contributions of inertia

CTR absolute contributions of inertia

## 6.5 Simulated study

### 6.5.1 Relative contributions to inertia (COR)

Average COR values for each missingness mechanism and across the six amounts of missingness are shown for each variable category in Figure 6.2.

COR values indicate the amount that each axis contributes to the inertia of the point. This makes it possible to identify the axis which contributes most to the inertia of each point. These values are scaled to add to 1000 across all dimensions. Because there are only two possible dimensions for this analysis, and axis 1 accounts for more than 80% of the total inertia, only the COR values for axis 1 are examined. Of the 15 variable categories, only one (ASMP) had a higher COR value on axis 2.

For the fully measured variable categories of 'age', 'gender' and 'area', no significant differences were found in COR values either across MM or for different M%. There were also no significant differences across for the asthma severity categories. However, significant decreases in COR values were found for ASNI and ASMP at 50% missingness.

Examining results for the variables with missingness, while the MM's did not show evidence of significant differences for the smoking category, SY, there were significant differences in the way these mechanisms behaved for SN. COR values for MNAR were significantly higher than for the other mechanisms and closer to the 'true' values. With regard to the amount of missingness, when compared to values at 5%, there was a significant reduction in the COR value for MNAR at 50% on SY and from 25% for MCAR on SN.

There were no significant differences across MM or M% for the NNY variable category. While no significant differences were found across MM for the NNN variable category, there was a significant drop in the COR values for MNAR from the 20% missingness point.

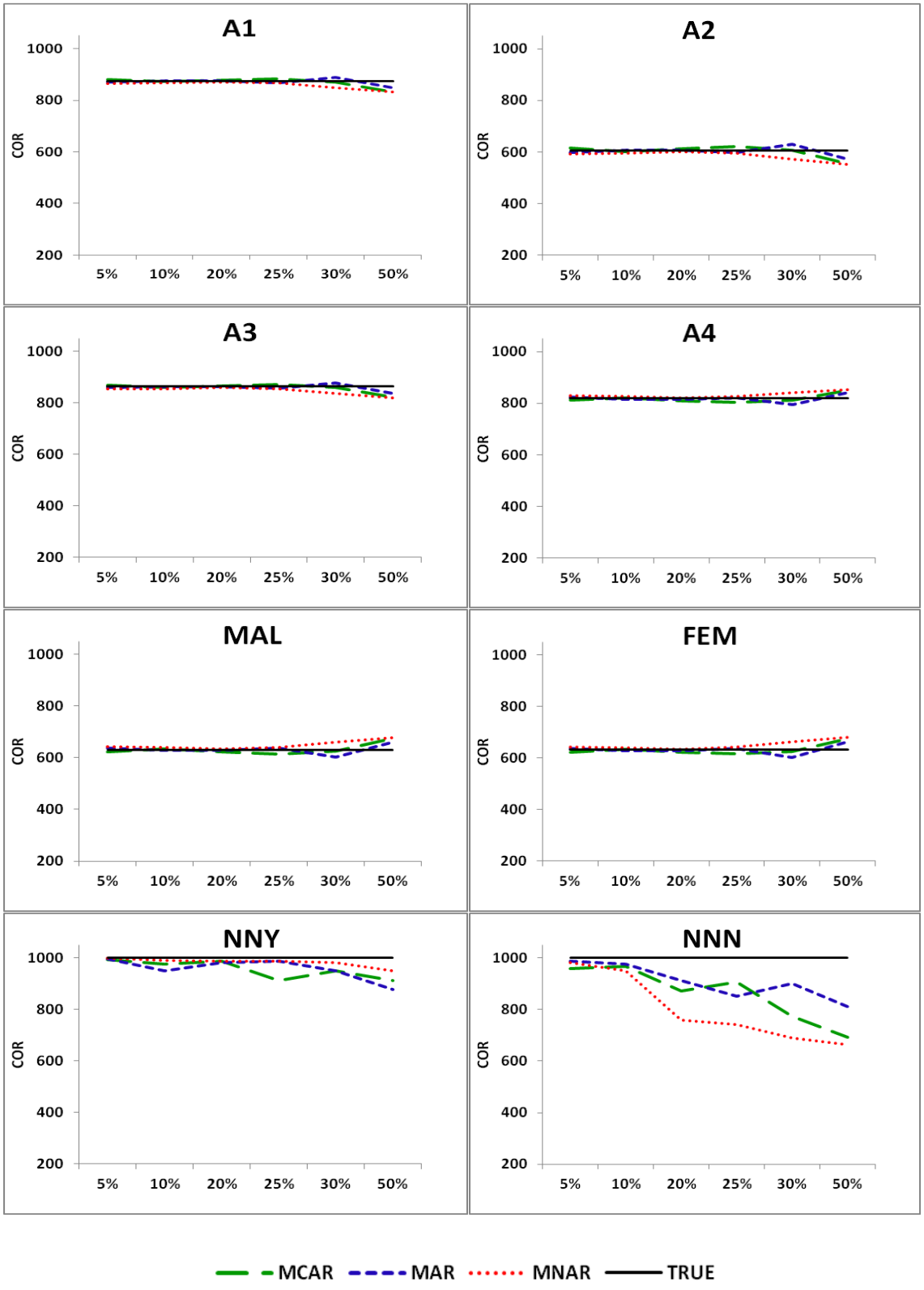



**Figure 6.2: COR values for each variable across all scenarios on axis 1**

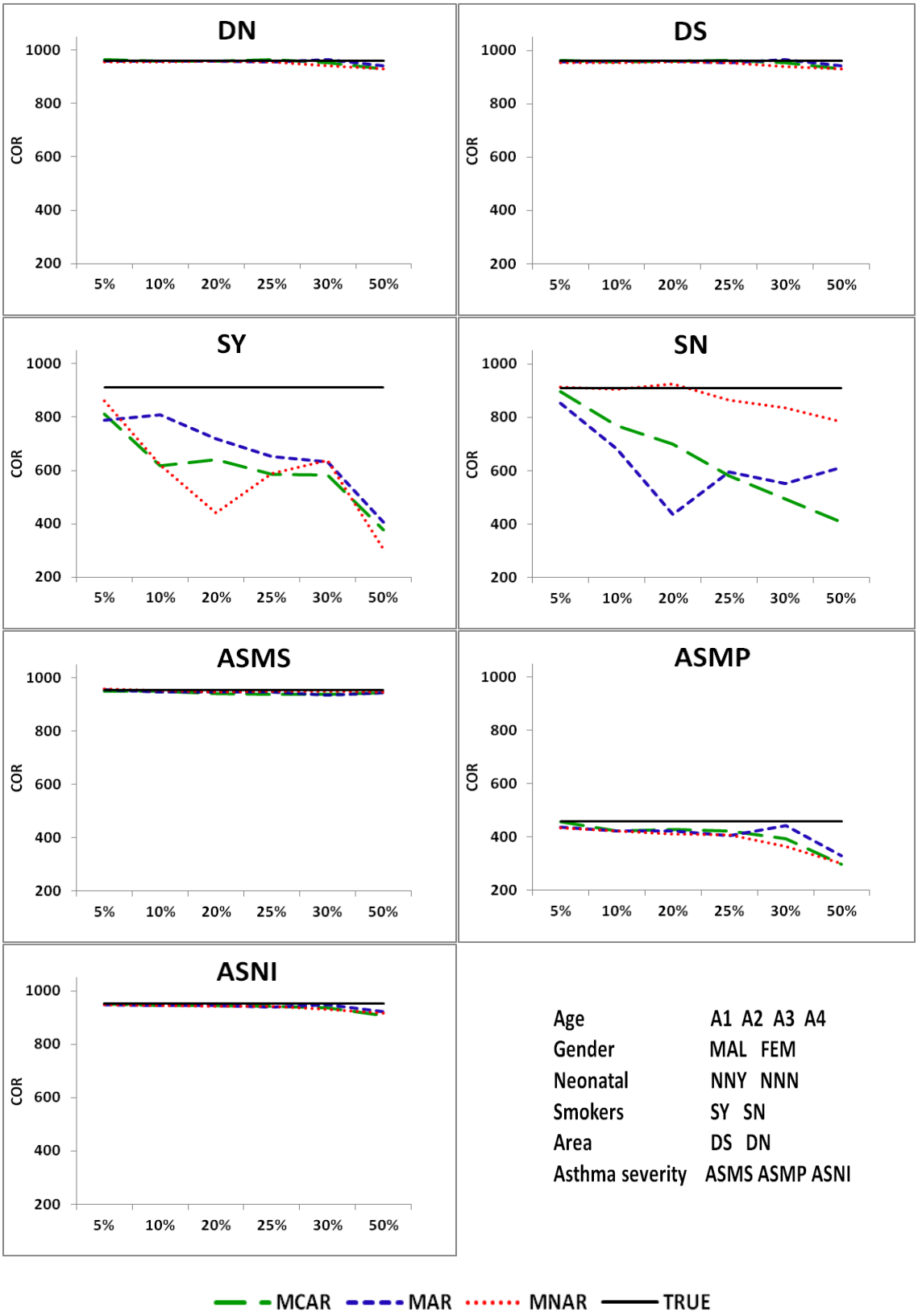



**Figure 6.2: COR values for each variable across all scenarios on axis 1 (continued)**

### 6.5.2 Absolute contributions to inertia (CTR)

Average CTR values for each MM and across the six M% are shown for each variable category in Figures 6.3 and 6.4.

CTR values, which have been scaled to sum to 1000 for each axis, indicate the amount that each variable contributes to the inertia of the axis. As a rule of thumb, if the CTR value of a point exceeds the average contributions of all the points (rows or columns) for a particular axis, then that point can be considered important to the orientation of the axis and is used in the interpretation of the results. In this study, an approximate CTR threshold value for the asthma severity categories is 333 and for the other variable categories it is 83.

**AXIS 1** – contrasts lowest asthma severity category (ASNI) with higher asthma severity categories (ASMP and ASMS)

Across all scenarios of MM and M%, A1, MAL, NNY and ASMS remained above the threshold value of importance while A2, A3, A4, NNN, SY, SN, DN, ASNI and ASMP remained below the threshold value of importance to this axis. Both FEM and DS were marginal with FEM positioned just above the threshold value and DS hovering around that value.

No significant differences between MM's were found for any variable category except NNN where, at 30% and 50%, CTR values for MAR were significantly higher than MNAR and MCAR values respectively. Furthermore, the only significant differences across M% were found for MAL, FEM and A4 where the CTR values increased significantly at 50% for all MM's.

**AXIS 2** – contrasts the two highest asthma severity categories: ASMS vs ASMP

All variable categories remained distinctly positioned relative to the threshold values except for SY which crossed the threshold at 50% for the MCAR and MNAR mechanisms. The only significant difference between MM's across all scenarios was at 25% on SN where the CTR value for MCAR is significantly higher than the value for MNAR. The only significant differences

across M%'s were found for MAL, FEM and A4 where CTR values decreased significantly at 50% for all MM's.

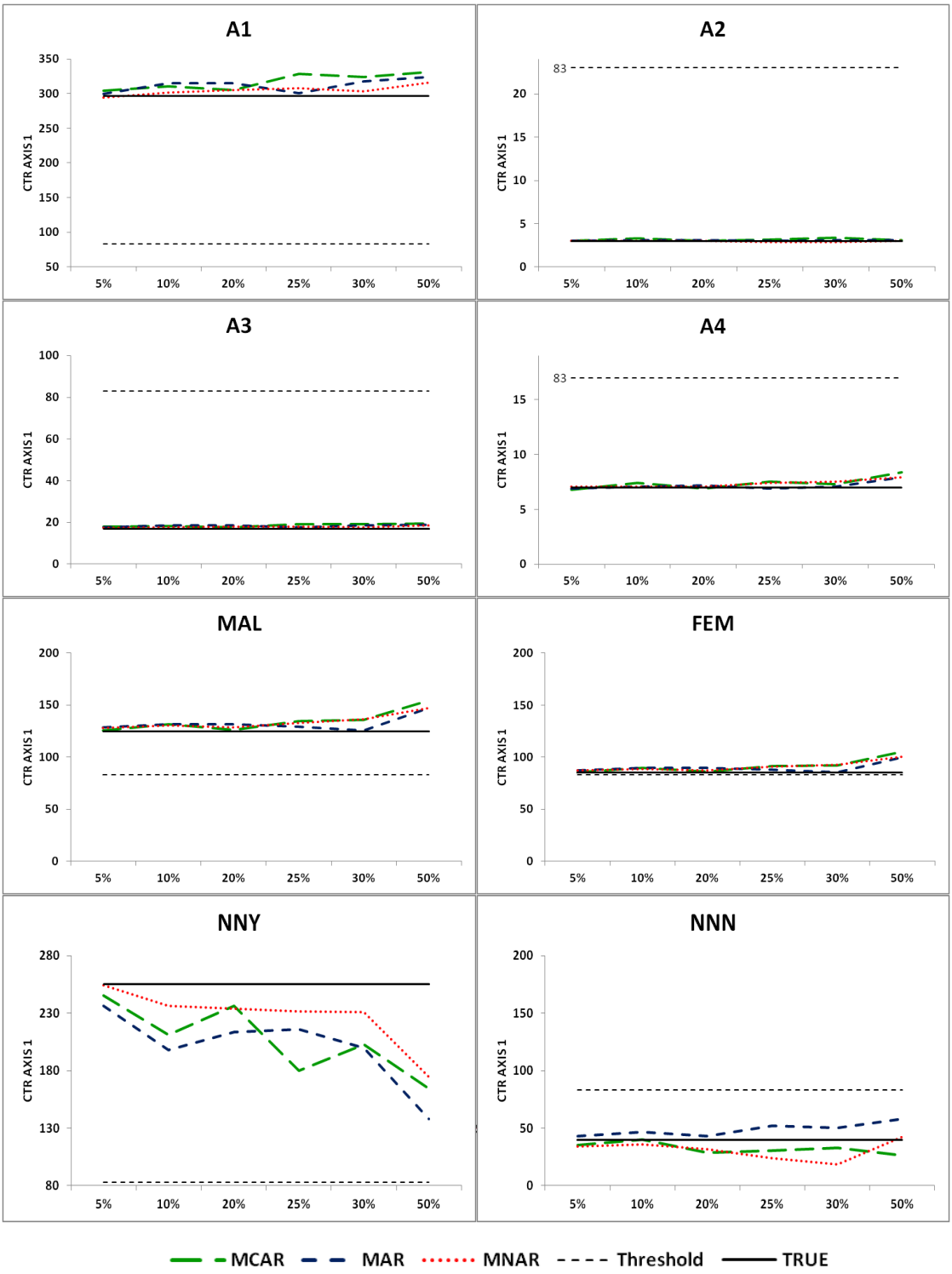



**Figure 6.3: CTR values for all variables on axis 1**

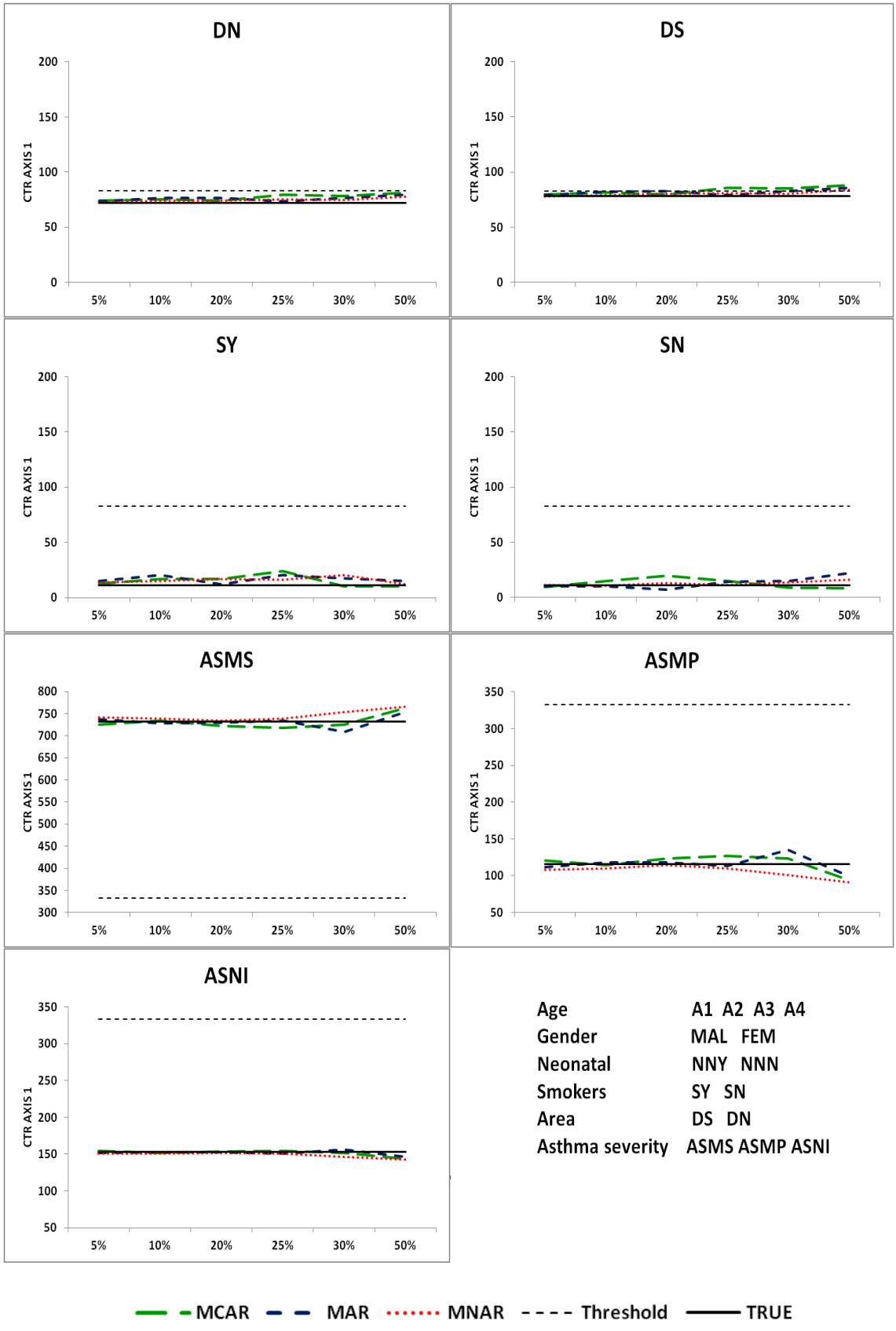



**Figure 6.3: CTR values for all variables on axis 1 (continued)**

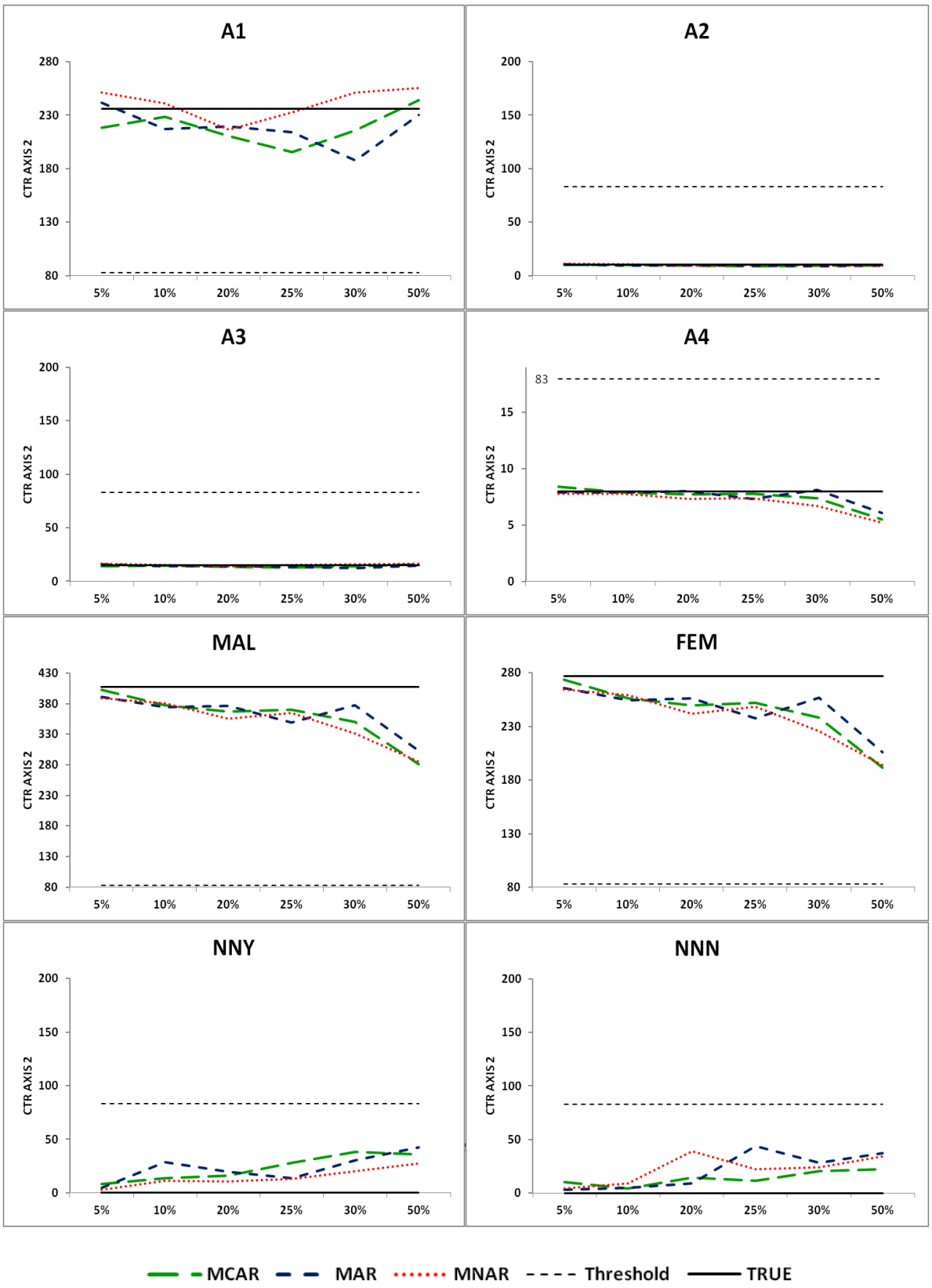



**Figure 6.4: CTR values for all variables on axis 2**

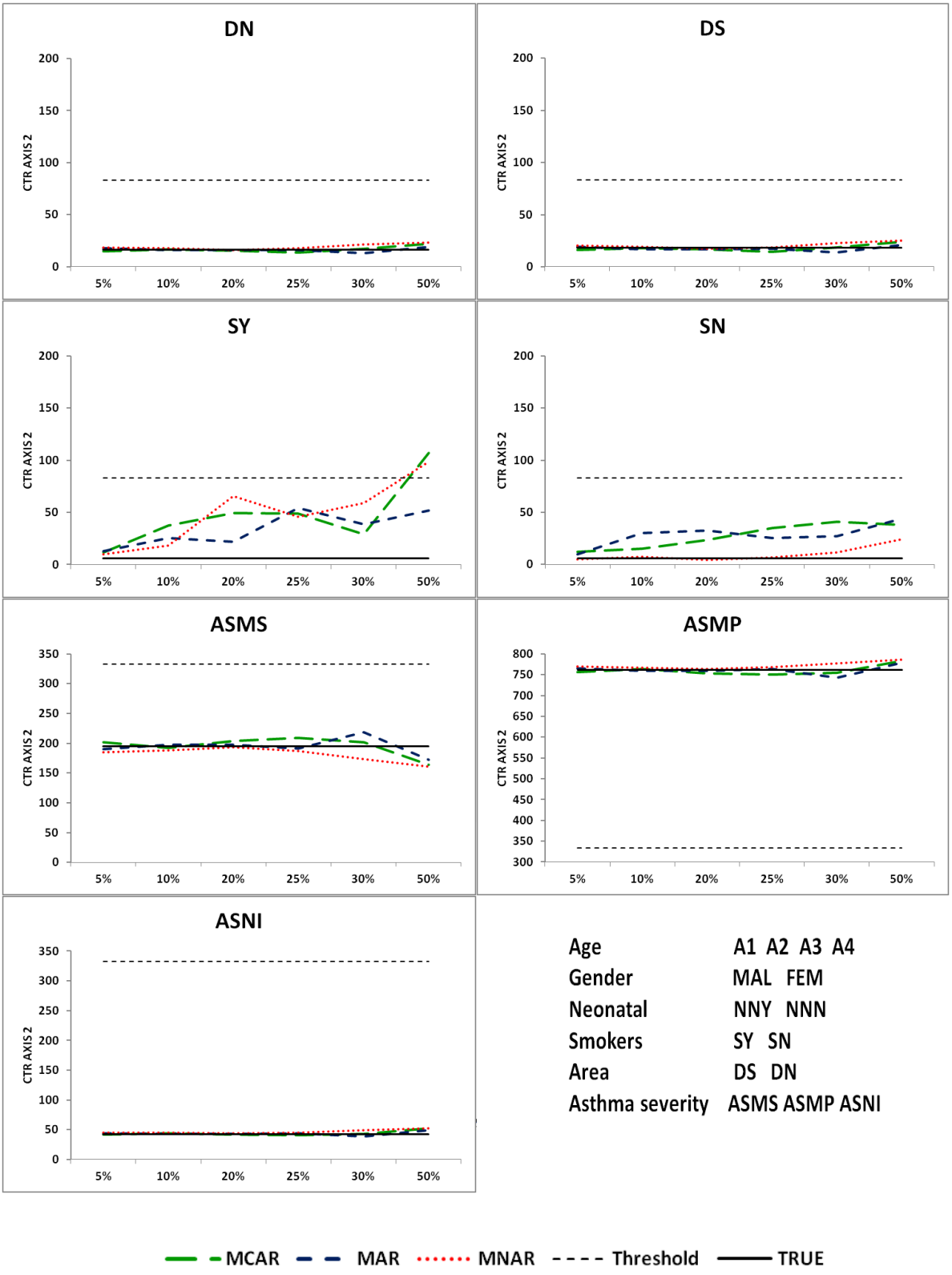



**Figure 6.4: CTR values for all variables on axis 2 (continued)**

### 6.5.3 Model inertia values

The total inertia of the measured data did not change significantly across either MM or M% (Figure 6.5(a)). However, the total inertia of the measured data, taken as a percentage of the total inertia of the full data set – measured and missing – was significantly higher for MCAR at 50% than at 5%. There was no significant difference in this measure across MM (Figure 6.5 (b)).

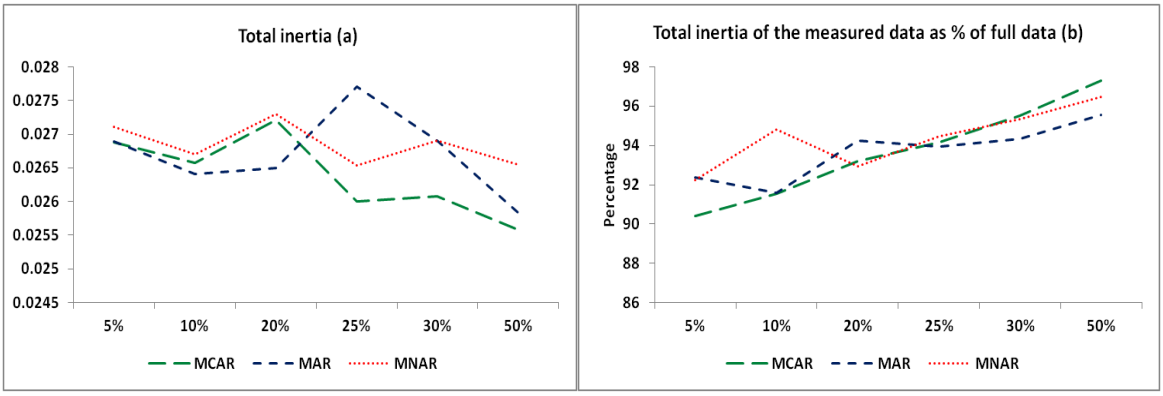



**Figure 6.5: Measures of inertia**

### 6.5.4 Graphical display

A degree of 'movement' is evident for some of the variable categories in the display of the subspace defined by axis 1 and axis 2 (Figure 6.6). This dispersion is more evident in the variables that have undergone missingness. For the variables that are fully measured, dispersion appears to be greater in those variable categories that have more variability and hence are situated further from the origin. On closer examination, it was found that variables further from their true positions are those with higher percentage missingness, with no specific correlation to mechanism.

## 6.6 Discussion

The aim of this simulation study was to explore the effect of both the missingness mechanism and the amount of missingness present in data on the use of s-CA as applied to categorical data that suffer from missingness.

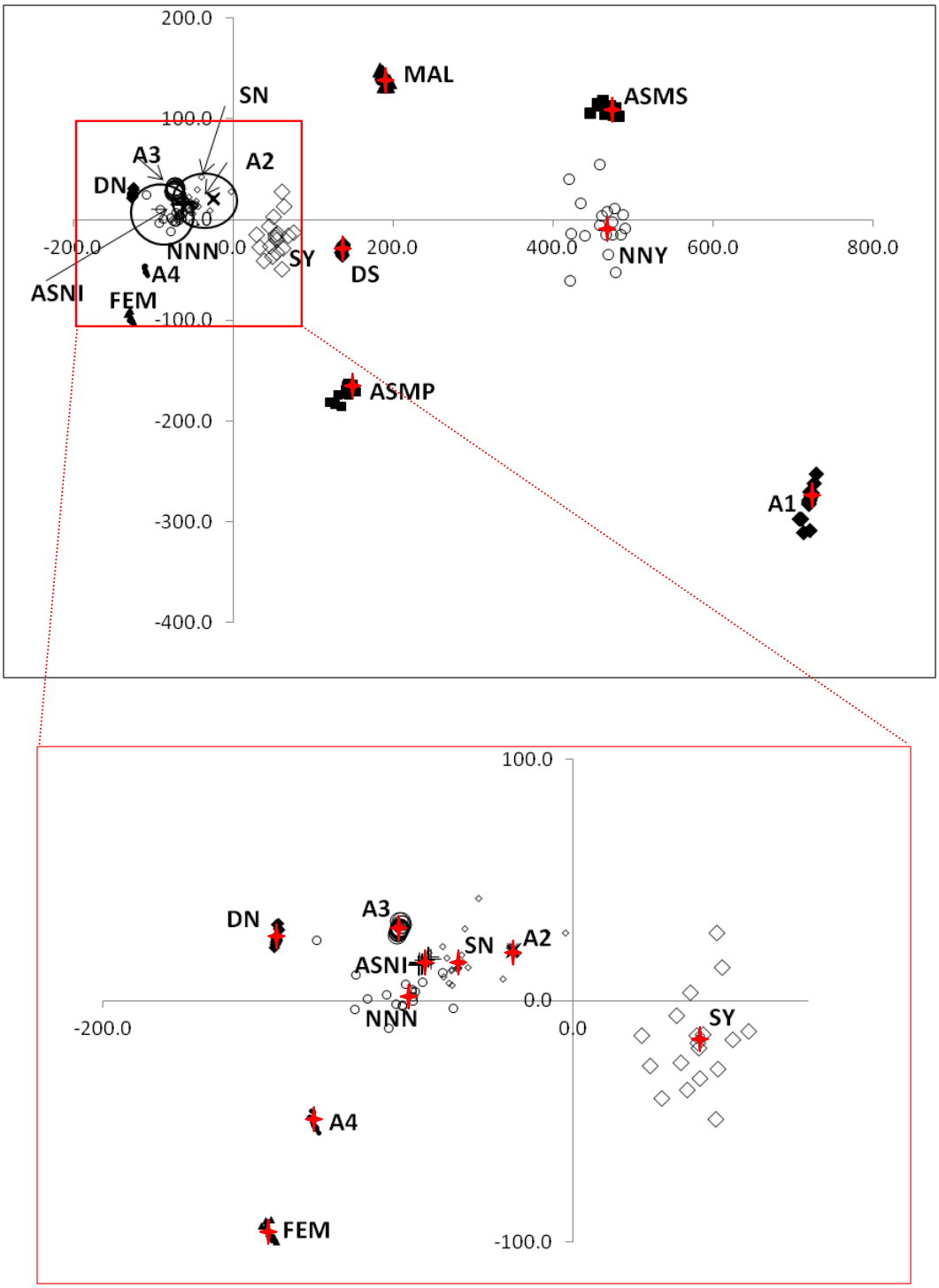
Three accepted MM's (MCAR, MAR and MNAR) were simulated, each across six degrees of missingness (5%, 10%, 20%, 25%, 30% and 50%). To allow for sampling variability, data for each of these 18 scenarios were generated 10 times and results were averaged.

It was found that the MM did not substantially affect the results. Furthermore, for missingness of up to 25% per variable, results were not notably affected.

The effect of the MM on all outcomes is negligible. In terms of the COR outcome, the only effect of the MM is on the variable category SN. In this case, MNAR values are significantly different from MCAR and MAR values but are in fact closer to the 'true' values.

Only one incident of significant difference across the MM's in CTR values for axis 1 was found. For the variable category NNN, CTR values for MAR are greater than those for MNAR and MCAR at 30% and 50% respectively. Likewise, the single significant difference across MM in CTR values for axis 2 was found for the variable category SN at 25% where the CTR value for MNAR is significantly smaller that the value for MCAR, but is in fact also closer to the 'true' value. Since the position of these measures relative to the threshold values remains the same, the differences do not affect the interpretation and these variable categories remain trivial to the orientation of axes 1 and 2.

These aforementioned differences across mechanisms all occur in the variable categories NNN and SN, which have been subjected to missingness but which are low in importance in terms of the orientation of the axes. The other variable categories that underwent missingness, SY and NNY, do not experience any significant differences across MM for any of the measures.




| Variable key: | | | | |
|---|---|---|---|---|
| | **Age** | A1/A2/A3/A4 | **Area** | DS/DN |
| | **Gender** | MAL/FEM | **Smokers** | SY/SN |
| | **Neo-natal** | NNY/NNN | **Asthma severity** | ASNI/ASMP/ASMS |
| | **True positions** | ✦ | | |

**Figure 6.6: Graphical representation of all variables for all scenarios.**

Greater deviations from the true COR values as well as variations in COR values across missingness mechanisms are apparent for the variable categories that underwent missingness. For each of these four variable categories, COR values are consistently lower than the 'true' values. In the case of SY and SN, COR values decrease from 910 to below 500 for some scenarios, thus indicating that axis 1 is no longer the most important axis to these points for all scenarios. This suggests that the importance of an axis to the inertia of a point decreases when missingness is introduced and can result in a change in the axis that contributes most to the inertia of a point. So the point can, at times, 'hop' axes. However, it is clear from the graphical display of the points that even when there is a drastic reduction in COR value, resulting in a possible 'hop' to another axis, the final placement of the point in the subspace is not compromised .

An effect of M% is found on COR values for the variable categories NNN, SY and SN. In the case of NNN, values of COR generally deviate more from the 'true' values as the percentage missing increases and there is a significant drop in COR value for MCAR at the 20% missing stage. A similar scenario exists for SN where there is a significant decrease in the COR value for MCAR as the amount of missingness increases. In the same way, the COR value for MNAR on SY is significantly lower at 50% missingness than at 5% missingness. The effect of M% on the asthma severity categories, ASNI and ASMP, indicates that there is a significant decrease in COR values at 50% missingness.

For those variables that did not undergo missingness, COR values across the three mechanisms do not differ appreciably from the true values for up to at least 25% missingness. Some differences, while not significant, are apparent for 30% and 50% missingness. Variables with missingness show erratic deviations from true values across all scenarios. These are especially pronounced in the variable categories that do not contribute appreciably to the general analysis.

The only significant effects of M% on CTR values for both axes are the significant change in values at 50% missingness for A4, MAL and FEM.

Apart from deviations from true CTR values being experienced by some incomplete variables, some deviation is also apparent in completely measured variables whose CTR values lie above the importance threshold. In no instance does this affect the overall outcome and interpretation.

While changes in the CTR values, in general, do not make a difference to the importance of points to the inertia of axes 1 and 2, there is one exception. CTR values for SY on axis 2 have increased from below to above the threshold level at 50% missingness for two of the MM's – MCAR and MNAR. This indicates that as missingness increases, it is possible for a variable to *become* 'significantly' important to the inertia of an axis. It must be remembered that, under the MNAR mechanism, data were deleted from the smoking variable at a ratio of 90:10 for SY:SN. Thus, compared to SN, a large proportion of data from SY would have been missing at 50% missingness. This would account for the greater effect of M% on SY than on SN under the MNAR mechanism and could indicate that under extreme missingness, analyses can lose some stability.

Total inertia for the measured data, a measure of the variability in the measured data, is not significantly affected by either the MM or the degree of missingness. However, when total inertia of the measured data is taken as a percentage of the total inertia of the full data set including the missing data, the percentage missing has an effect. For MCAR, this measure at 50% missingness is significantly higher than at 5% missingness. Visually, there is an upward trend across all mechanisms as missingness increases. This may imply that the variability in the missing data is significantly lower for the MCAR mechanism at 50% missingness than at 5% missingness.

Examining the plot of all variables across the 18 scenarios confirms that there is some deviation from the true position for some variables. This is most evident in the variables that suffer from missingness; but is also present in the completely measured variables that show stronger associations with asthma severity (MAL and A1). In general it was found that, while points that are further from their 'true' position have a higher percentage, there is no evidence that the MM is a factor in this displacement. In all cases, the dispersion is well contained and the relative positioning of variable categories with each other and with the axes remains unchanged.

Results from this study suggest that there is no evidence that the type of MM has an effect on results after applying s-CA to data that suffer from missingness. In addition, the analysis has shown that only when missingness exceeds the 30% level, are some results affected. However, while deviations in the outcomes studied are present, they do not affect the overall interpretation of the analysis.

### 6.6.1 Limitations

While the results emanating from this study are reliable, some limitations have been noted. These results are specific to the variables included and the mechanisms imposed on this data. The variables were selected according to their relationships with asthma severity such that all strengths of relationship are represented. Three of the four variable categories that underwent missingness do not have strong relationships with asthma severity. For ease of interpretation, missingness was imposed on only two of the six variables. Furthermore, while the deletions on these variables were based on plausible judgements, they are subjective, and may have influenced the findings. Further studies need to be carried out to explore the effect of different ratios of missingness and a different or increased choice of variables.

Ten simulations were performed on this data. The number of simulations to perform is dependent on the required accuracy with an increase in the number of simulations resulting in more accuracy (Burton et al., 2006, Ritter et al., 2011). In contrast to confirmatory techniques, in which relationships are hypothesised and proved, CA (and its variants) is an exploratory approach in which relationships in the data are revealed and visualized for purposes of interpretation. Relative positions of category points indicate levels of similarity or association between categories. No measures of statistical significance are applied (Greenacre, 1992). Thus accuracy is not of prime importance. There is little evidence to suggest that additional simulations would have produced meaningfully different results.

## 6.7 Conclusions

Under the conditions imposed in this study, no evidence was found to suggest that the missingness mechanism has an effect on results when s-CA is applied to data that suffer from

missingness. It was found that, in some cases, values of the outcomes studied deviate from the true values when the amount of missingness exceeds 30% per variable. These deviations do not, however, affect the overall interpretation of the results. It is believed that s-CA would have a similar impact on other data sets that comprise categorical variables that suffer from missingness.

# Chapter 7

# A COMPARATIVE STUDY OF MULTIPLE IMPUTATION AND SUBSET CORRESPONDENCE ANALYSIS

The applications of MI and s-CA to manage missing data were presented in Chapters 4 and 5 respectively. In this chapter, interactions are introduced to analysis with s-CA and a comparison of the two methods is made. In order to compare the outcomes of these two diverse methods, a set of core variables and interactions is chosen to be used in the application of both methods. Results are compared and similarities and differences in the methods and the outcomes they produce are highlighted.

## 7.1 Introduction

Both MI and s-CA have successfully been used to identify associations between asthma severity and several environmental, socio-economic, genetic and behavioural variables in which missingness is present (Hendry et al., 2014a, Hendry et al., 2014b). It is also evident from the details of each method in Chapter 3 that the methodologies adopted by these two methods are vastly different.

MI works under the assumptions of distributions and missingness mechanisms and fits the data to a pre-assumed model. In contrast, s-CA has no restrictions with respect to distributions and missingness mechanisms. The only requirement for application of this method is that the data be non-negative and categorical – a condition easily achieved through simple transformations. Thus no model is assumed but rather the data are decomposed to reveal their trends and relationships.

The two approaches, therefore, have no common parameter estimates or model structure. Thus the conventional comparison of methods in terms of mean square errors or goodness of fit is not directly applicable. Accordingly a systematic holistic review of the two approaches is

adopted. Since results in this study from MVNI and FCS were shown to be similar (Chapter 4), only analysis using the FCS imputation algorithm is included in this comparative study as it is more suited to categorical data than MVNI.

## 7.2 Selected variables

The 17 variables identified for inclusion in the analysis with MI, following model building (Chapter 4), were selected to be used in this comparative study. For the analysis with s-CA, the interval variable 'age' was transformed into a 4-level categorical variable, as before. For the purpose of comparison, the 3-tiered asthma severity variable was used for both analyses. This data has been previously detailed (Chapters 2, 4 and 5)

Of the 10 identified interactions, the two strongest – 'gender * smoke exposure in vehicles' and 'fear * breakfast habits' were included in this analysis.

## 7.3 Adding interactions to subset correspondence analysis

In order to incorporate the interactions into a s-CA, each interaction was broken down into its product terms which were then treated as additional categories in the data.

For example, the interaction gender (male/female) * smoke exposure in vehicles (yes/no) was broken down into: male/yes; male/no; female/yes and female/no. These categories were then added to the contingency table as extra columns.

Variables involved in the interactions – 'gender', ' smoke exposure in vehicles', 'fear' and 'breakfast habits' – were not included as individual active variables in the analysis but were treated as supplementary variables (Greenacre, 1984). By so doing, they do not participate in the orientation of the axes but their individual positions as 'main effects' relative to the associated interactions (Torres-Lacomba, 2006) can still be studied.

## 7.4 Results

The aim of this investigation was to illustrate and compare two methods to analyse categorical data that suffer from missingness. It was found that, while MI, in combination with ordinal regression, and CA applied to a subset of data are vastly different methodologically, the results that they produce in the analysis of inter-variable relationships are very similar.

The application of these two methods enabled the identification of relationships between asthma severity and several environmental, genetic, socio-economic and behavioural variables while, at the same time, retaining all records. Furthermore, the associations between these variables and asthma (Table 7.1 and Table 7.2) were consistent across methods and generally confirmed established theories regarding factors that exacerbate asthma. There was agreement that confirmed asthma is associated with children who: are younger (Asher et al., 2006); have had some special neo-natal care (Mai et al., 2003); are exposed to smoke in the home (Charoenca et al., 2013, Ehrlich et al., 1992), in vehicles (Sendzik et al., 2009), *in utero* (DiFranza et al., 2004) and in the form of air pollution (Neidell, 2004, Peden, 2005); lived in a home with up to 4 people (Jarvis et al., 1997); come from a R4501 – R10000 income household; do not always have enough food; are exposed to low concentrations of compounds and pollutants (Becher et al., 1996, Venables and Chan-Yeung, 1997); never had a pet and do not experience fear in the neighbourhood. Both analyses also indicated an association between worse asthma and both lack of violence in the neighbourhood and watching up to one hour of TV a day. These associations are contrary to what other studies have found and, while the data were explored for reasons for these anomalies, none were found. It was concluded that there must be some underlying factor specific to this sample.

The interpretation of the interactions was also consistent across methods. With regard to the 'gender * smoke exposure in vehicle' interaction, it was found that male children who are exposed to smoke in a vehicle suffer from significantly worse asthma than girls not exposed to smoke in a vehicle. Further, amongst the females in this study, those who are exposed to smoke in a vehicle suffer from less severe asthma than those not exposed to smoke in a vehicle. For those in the study not exposed to smoke in a vehicle, asthma severity is marginally worse for the males.

**Table 7.1: Estimated coefficients (EST) and standard errors (SE)**

| Predictor | Reference Category | Category | FCS(N = 382) | |
|---|---|---|---|---|
| | | | EST | SE |
| Gender | Female | Male | 0.039 | 0.362 |
| Neo-natal care | No | Yes | 0.847* | 0.394 |
| Fear | No | Yes | -1.042* | 0.406 |
| Smoked while pregnant | No | Yes | 0.379 | 0.488 |
| Smokers in home | No | Yes | 0.701* | 0.309 |
| Smoke in vehicles | No | Yes | -0.706 | 0.512 |
| Exercise | >4 times a week | Up to once a week | 0.044 | 0.384 |
| | | 2 – 4 times a week | 0.011 | 0.384 |
| TV watching | >3 hours a day | Up to 1 hour a day | 0.786 | 0.465 |
| | | 1 – 3 hours a day | 0.046 | 0.43 |
| Number people in home | 8+ | 1 - 4 | 0.981* | 0.481 |
| | | 5 - 7 | 0.381 | 0.494 |
| Income | R100001+ | up to R1000 | -0.133 | 0.611 |
| | | R1001 – R4500 | 0.017 | 0.555 |
| | | R4501 – R10000 | 0.697 | 0.523 |
| Food availability | Enough | Not always enough | 0.756 | 0.406 |
| Work'nWear | No | Yes | 0.402 | 0.456 |
| Pets ever | No | Yes | -1.072* | 0.398 |
| Area | North Durban | South Durban | 0.595 | 0.306 |
| Breakfast habits | Daily | Not daily | -1.011* | 0.494 |
| Violence | No | Yes | -0.709* | 0.34 |
| Age | | | -0.247 | 0.16 |
| Fear * Breakfast | No/Daily | Yes/Not daily | 2.338* | 0.725 |
| Gender * SmokeVehicle | Female/No | Male/Yes | 1.811* | 0.699 |

FCS -Multiple imputed FCS

Interpretation of the 'fear * breakfast habits' interaction showed that, compared to those who do not experience fear and eat breakfast daily, there is a significant chance that those who do experience fear but do not eat breakfast daily will suffer from worse asthma. Results also indicate that for those who eat breakfast daily, worse asthma is experienced by those who do not experience fear than by those who do experience fear. Furthermore, for those who do not experience fear, children who eat breakfast daily have marginally worse asthma than those who don't eat breakfast daily. Whereas with s-CA the classifications as supplementary

variables of those variables included in the interactions enabled the study of their positions relative to the asthma severity categories (male children suffer from worse asthma than female children (Almqvist et al., 2007, Bonner, 1984)), this was not possible with the MI approach.

While on the surface these methods produce the same overall results, a deeper study of the results identified several differences in the outcomes from these methods.

Whereas with MI and ordinal regression the interpretation of results indicated the relative severity of asthma from one category to another category of a specific variable, with the application of CA the identification of factors associated with the specific asthma severity classifications was possible. To illustrate this point, analysis with MI and ordinal regression showed that worse asthma is experienced by those who had neo-natal care than by those who did not have any neo-natal care. On the other hand, results from s-CA were more specific and having had neo-natal care was shown to be associated with moderate to severe asthma while not having neo-natal care is associated with mild intermittent or no asthma.

An appealing feature of the MI approach is the ability to calculate the odds of severity of one category of asthma severity relative to the complementary categories for various explanatory variables relative to a reference category. It is seen, for example, that children who have had neonatal care are more than twice as likely (OR = $e^{.847}$ = 2.22) as those who have not had neonatal care to suffer from worse asthma. Other significant results show that: children who come from a home where people smoke are twice as likely to suffer from worse asthma (OR = 2.02) compared to those who don't live with smokers; and those who come from a home with up to four people are 2.67 times as likely to experience worse asthma (OR = 2.67) than those in a home with eight or more people. Compared to children with no pets, the odds that children with pets will suffer from worse asthma are approximately one third. Children who experience fear in their neighbourhoods and do not eat breakfast daily are more than 10 times as likely to suffer from higher levels of asthma severity as children who do not experience fear and eat breakfast daily. It was also found that the odds of suffering from worse asthma are six times higher for boys who travel in vehicles with smokers than for girls who do not travel in cars with smokers.

**Table 7.2: Decomposition of inertia for the 2 principal axes**

| Name | Mass | INR | k= 1 | COR | CTR | k= 2 | COR | CTR |
|---|---|---|---|---|---|---|---|---|
| A1 | 4 | 3 | -717 | 955 | 146 | -156 | 45 | 42 |
| A2 | 34 | 0 | 34 | 754 | 3 | 19 | 246 | 5 |
| A3 | 24 | 0 | 73 | 874 | 8 | 28 | 126 | 7 |
| A4 | 5 | 0 | 58 | 151 | 1 | -138 | 849 | 34 |
| NNY | 9 | 3 | -476 | 985 | 129 | 59 | 15 | 12 |
| NNN | 55 | 0 | 66 | 1000 | 16 | -1 | 0 | 0 |
| SPY | 6 | 3 | 13 | 27 | 0 | -75 | 973 | 14 |
| SPN | 57 | 57 | -7 | 454 | 0 | 7 | 546 | 1 |
| SY | 33 | 23 | -51 | 969 | 6 | -9 | 31 | 1 |
| SN | 34 | 34 | 47 | 972 | 5 | 8 | 28 | 1 |
| E1 | 20 | 11 | -2 | 480 | 0 | -2 | 520 | 0 |
| E2 | 24 | 6 | 60 | 667 | 6 | -42 | 333 | 17 |
| E3 | 19 | 24 | -14 | 62 | 0 | 56 | 938 | 24 |
| T1 | 15 | 2 | -186 | 454 | 34 | 204 | 546 | 248 |
| T2 | 34 | 4 | 45 | 326 | 4 | -65 | 674 | 56 |
| T3 | 14 | 3 | 147 | 997 | 19 | -8 | 3 | 0 |
| N1 | 22 | 18 | -170 | 994 | 41 | -13 | 6 | 1 |
| N2 | 27 | 7 | 61 | 983 | 7 | 8 | 17 | 1 |
| N3 | 12 | 48 | 160 | 934 | 21 | 43 | 66 | 9 |
| I1 | 14 | 0 | 41 | 952 | 2 | 9 | 48 | 0 |
| I2 | 18 | 13 | 33 | 501 | 1 | -33 | 499 | 8 |
| I3 | 15 | 4 | -191 | 992 | 37 | 18 | 8 | 2 |
| I4 | 7 | 73 | 102 | 977 | 5 | 16 | 23 | 1 |
| Fn | 46 | 31 | 59 | 938 | 11 | 15 | 62 | 4 |
| Fe | 15 | 6 | -31 | 435 | 1 | 35 | 565 | 7 |
| WWY | 6 | 11 | -406 | 915 | 68 | -124 | 85 | 38 |
| WWN | 58 | 35 | 35 | 736 | 5 | 21 | 264 | 10 |
| PY | 20 | 2 | 199 | 951 | 51 | 45 | 49 | 16 |
| PN | 46 | 42 | -80 | 998 | 19 | -3 | 2 | 0 |
| DS | 33 | 15 | -125 | 997 | 33 | -6 | 3 | 1 |
| DN | 34 | 6 | 120 | 997 | 32 | 6 | 3 | 1 |
| VY | 32 | 12 | 111 | 991 | 26 | 10 | 9 | 1 |
| VN | 29 | 25 | -95 | 978 | 17 | -14 | 22 | 2 |

Mass (Mass) and inertia (INR) of each variable; the quality (QLT) of the variable's representation in the subspace of the first 2 axes; Co-ordinates (k = ...); contributions of axes to the inertia of the variables (COR); and contributions of variables to the inertia of the axes (CTR)*.

**Table 7.2: Decomposition of inertia for the 2 principal axes (continued)**

| Name | Mass | INR | k= 1 | COR | CTR | k= 2 | COR | CTR |
|---|---|---|---|---|---|---|---|---|
| FYBd | 20 | 12 | 179 | 987 | 41 | 20 | 13 | 3 |
| FNBd | 21 | 78 | -203 | 931 | 58 | -55 | 69 | 26 |
| FYBn | 9 | 6 | -141 | 654 | 12 | 103 | 346 | 38 |
| rNBn | 12 | 9 | 228 | 957 | 40 | 48 | 43 | 11 |
| mSVY | 7 | 276 | -363 | 998 | 60 | 16 | 2 | 1 |
| mSVN | 19 | 22 | -76 | 158 | 7 | 176 | 842 | 228 |
| fSVY | 9 | 74 | 137 | 947 | 12 | -32 | 53 | 4 |
| fSVN | 27 | 3 | 104 | 480 | 19 | -109 | 520 | 124 |
| | | | | | | | | |
| ASMS | 71 | 89 | -371 | 929 | 635 | 103 | 71 | 295 |
| ASMP | 123 | 130 | -157 | 638 | 198 | -118 | 362 | 679 |
| ASNI | 806 | 781 | 56 | 975 | 168 | 9 | 25 | 26 |
| **SUPPLEMENTARY** | | | | | | | | |
| MAL | | | -150 | 516 | | 146 | 484 | |
| FEM | | | 107 | 494 | | -108 | 506 | |
| FrY | | | 66 | 808 | | 32 | 192 | |
| FrN | | | -50 | 896 | | 17 | 104 | |
| SVY | | | -79 | 980 | | -11 | 20 | |
| SVN | | | 27 | 894 | | 9 | 106 | |
| Bnd | | | 68 | 462 | | 73 | 538 | |
| Bd | | | -22 | 596 | | -18 | 404 | |

Mass (Mass) and inertia (INR) of each variable; the quality (QLT) of the variable's representation in the subspace of the first 2 axes; Co-ordinates (k = ...); contributions of axes to the inertia of the variables (COR); and contributions of variables to the inertia of the axes (CTR)*.

Compared to the MI approach, CA was also able to identify specific associations that distinguished between the different levels of variables. This can be seen with the 'TV watching' variable. Results from the MI approach indicate that the amount of TV watched is inversely related to the severity of the asthma and in fact children from this study who watch up to one hour of TV a day are more than twice as likely to suffer from worse levels of asthma as those who watch in excess of 3 hours a day. With s-CA, by examining the positions of these variable categories in the graphical display (Figure 7.1), as well as the decomposition of the inertia (Table 7.2), it can be seen that watching 1 hour of TV a day (T1) is associated with moderate to severe asthma; watching between 1 and 3 hours a day (T2) is associated with mild persistent

asthma; and watching more than 3 hours a day (T3) is associated with mild intermittent or no asthma. Thus a finer distinction is possible regarding categories of variables and their associations with levels of asthma severity.

By using the graphical display produced by s-CA, it is possible to identify inter-variable relationships that do not include the asthma severity variable. For example, the positions of the variables I3, N1, DS and VN indicate that they share some relationship. This is not possible with the MI approach, when considering categorical data.

With CA, it was also possible to compare the strengths of association with asthma severity of several predictor variables. For instance, from the positioning of the points on the display, it can be deduced that while the risk of having moderate to severe asthma from smoke exposure in a vehicle exceeds the risk from smoke exposure in the home (Sly et al., 2007) or smoke exposure in utero, the greatest risk is from air pollution as experienced in the South Durban region.

All these factors discussed above illustrate the extent of the usefulness of the graphical display produced by s-CA as a tool to identify inter-variable relationships.

Unlike the analysis with MI and ordinal regression, inter-variable relationships found to exist with the application of CA cannot be assumed to be statistically significant. While the relative strength of associations can be deduced by examining the angles that the points made with each other and with the principal axes in the graphical display (Hendry et al., 2014a), these results cannot be projected onto a broader population. Results from s-CA indicated that theassociation of ASMS (moderate to severe asthma) with NNY (having had neo-natal care) is stronger than its association with DS (South Durban) as seen by the relative size of the angles between them. These results are confirmed in the MI and ordinal regression analysis and, in addition, the significance of the association between neo-natal care and asthma severity is indicated.

Because MI is computationally intensive, complications and limitations can be encountered. This can occur with large data sets and even more so when a large number of variables suffer

from missingness (Lee and Carlin, 2010, Van Buuren, 2007). The need to include many interactions in the imputation model in order to ensure that it is more general than the analysis model, is often not feasible and computationally not possible (Lee and Carlin, 2010) – especially with data sets that have a large number of variables. These problems were not encountered with this analysis despite the seemingly large number of variables. In fact, in a previous study using this data (Hendry et al., 2014b), 10 interactions were included with no problems being experienced. In contrast, computationally, CA can cope with large numbers of variables and interactions, but this can cause overcrowding in the display which makes it difficult to identify points and interpret relationships between them. It is for this reason that the number of interactions in this study was limited to two. The possibility does, however, exist with s-CA to include more interactions and analyse them as a separate subset.

Preliminary analysis of this data set indicated that the missingness is at best MAR with a possibility of some MNAR present (Hendry et al., 2014b). Because MI produces unbiased estimates providing that the missingness is at worst MAR, it was necessary to include, in the imputation model, variables associated with the missingness of the incomplete variables, the outcome variable – asthma severity – as well as the two interactions chosen for the analysis model. This inclusion of carefully selected variables should produce acceptable results even if some MNAR is present (Graham et al., 1997). In contrast to this, CA and its variants are not constrained by complexities of models or distribution requirements. It is also not sensitive to the missingness mechanism in the data (Hendry et al., 2015). Therefore no special adjustments were needed to counteract the possibility of some MNAR missingness. The only adjustment needed in this study was to categorise the interval variable ‘age’. While non-negative categorical data is a requirement of CA, it is generally a straightforward exercise to achieve this condition.

The fact that only a few of the variables in the MI/ordinal regression analysis were significantly associated with asthma severity is consistent with the results from s-CA. The visible bunching up of the points in the graphical display and the low inertia values – a total of only 0.0178 – indicate that only a limited amount of variability is present in this data (Greenacre, 1992).

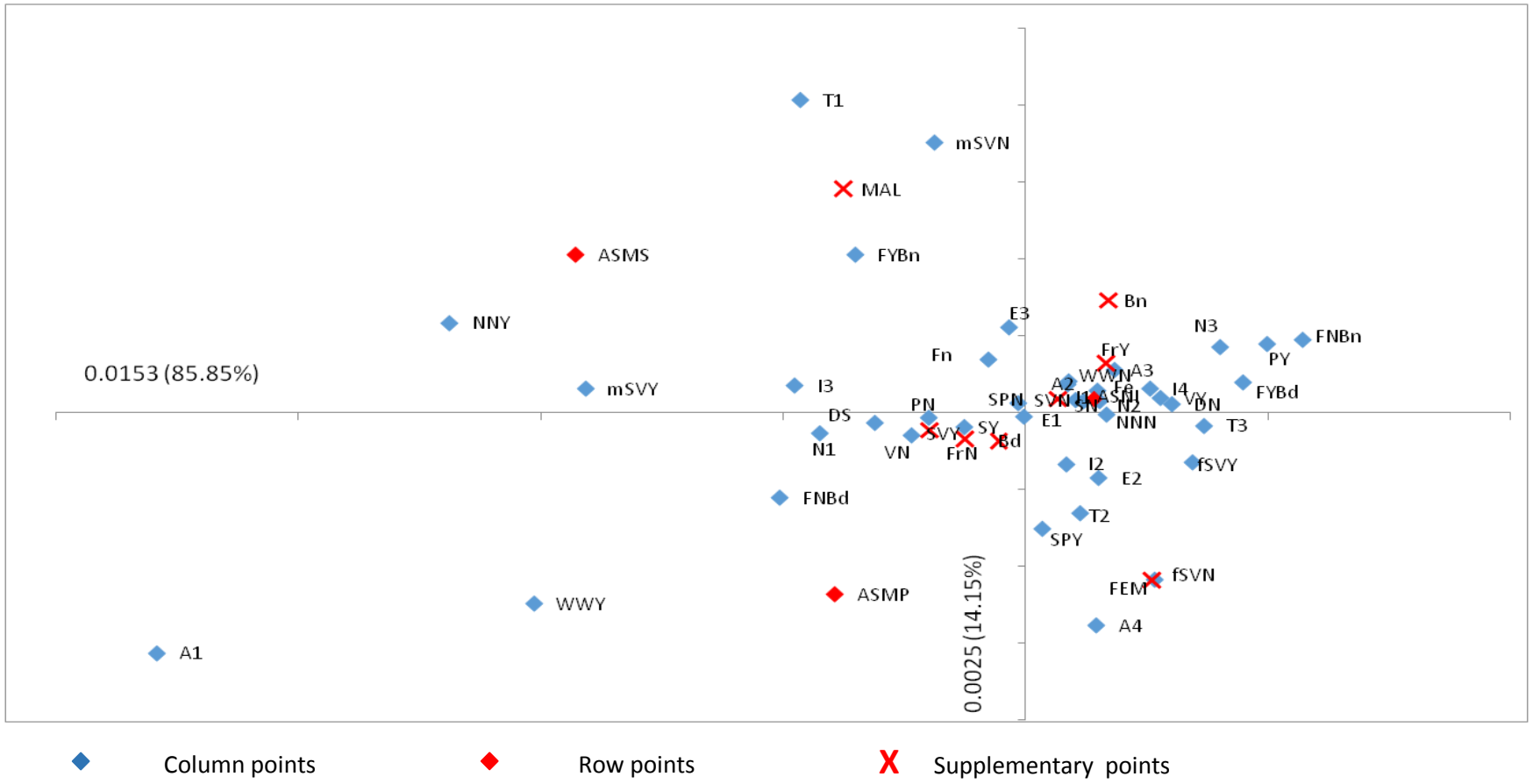



**Figure 7.1: s-CA map of a contingency table with the row points and column points projected onto the plane of the first and second principal axes. Values on the axes indicate principal inertias and their respective percentages of total inertia.**

## 7.5 Conclusions

Non-response is a reality in survey data and needs to be handled appropriately. The use of MI in conjunction with ordinal regression as well as CA as applied to the subset of measured data to analyse categorical data that suffer from missingness has been demonstrated. The addition of interactions to an analysis with s-CA has also been demonstrated. General relationships between the environmental, socio-economic, genetic and behavioural variables and asthma severity were found to be consistent across methods. Each method offers a different set of advantages in their applications. Analysis with s-CA is less demanding than with the MI approach – both in terms of conditions and the computational process – and finer distinctions in the inter-variable relationships can be made. These relationships are, however, 'looser' than those obtained from the MI approach and significance cannot be claimed. Despite their differences, the results produced in this investigation provide support for the greater use of less restrictive and less computationally intensive graphical methods to analyse categorical data that suffer from missingness.

# Chapter 8

# Conclusions

This study has striven to address the issue of missing categorical data by investigating the use of two diverse methods as applied to a set of data gathered from children with and without asthma in Durban, South Africa.

Two approaches to MI, MVNI and FCS were presented and their results compared. Specific emphasis was placed on the inclusion of interactions not known *a priori*. In contrast to MI, the exploratory graphical technique of subset correspondence analysis (s-CA) was illustrated as a tool to manage missing categorical data. Furthermore, an investigation was made into the effect of the mechanism and amount of missingness present, on outcomes of s-CA.

While imputation with MVNI assumes normality of variables and imputes all missing values according to the multinomial normal distribution, FCS is more flexible and imputes values in accordance with the type of variable suffering from missingness. This study showed that results from these two approaches, when applied to predominantly categorical data, were very similar, with the size and direction of association between all predictor variables and the outcome variable, asthma severity, remaining consistent across methods. There were some minor discrepancies in significance levels but these only affected variables with more than two categories. It is thought that these discrepancies could be as a result of the imputation model used and therefore dependent on the assumed distribution of the imputation model. This study also investigated and presented a method to identify interactions to be included in the imputation model by making use of a single data set imputed from the ML estimates of the EM algorithm for covariance matrices. While overall results produced from the two methods did not differ substantially, FCS was easier to apply to the categorical data in that no adjustments to the data were necessary either before or after imputation. This would suggest that FCS is a more suitable choice of imputation method for analysis of missing data in which variables are primarily categorical.

Where the more traditional methods of MI in conjunction with some regression analysis make assumptions on distributions and fit the data to a model, CA and its variants, including s-CA, have no such restrictions. While data are required to be categorical and non-negative, these conditions can usually be realized with some minor transformations prior to analysis. This makes this family of graphical multivariate techniques simple to use and applicable to most situations.

This study illustrated the application of s-CA and successfully identified relationships between several factors and asthma severity without having to deal with the complications of the missing data. The missing data were managed by introducing separate 'missing' categories for each variable that suffered from missingness. Furthermore, the overcrowding of the graphical display and domination by missing categories that can occur and make the identification of relevant inter-variable associations difficult, was alleviated by the analysis of the subset of data that excluded all 'missing' categories. This study also illustrated the addition and interpretation of interactions as applied to s-CA. A simulation study investigating the effect of the missingness mechanisms on outcomes of s-CA suggested that this method is not sensitive to the missingness mechanism present in the data. It was further found that, even though missingness of more than 30% affected some of the outcomes, the overall interpretation of relationships was not compromised. The results of this study suggest that s-CA is a suitable method to use when investigating relationships between categorical variables that are affected by non-response.

In comparing these two aforementioned methods to identify inter-variable relationships in the presence of missing data, several differences were apparent in their application requirements and assumptions and in their outcomes.

Because the presence of MNAR missingness in this data could not be ruled out, special care was needed in the application of MI to include, in the imputation model, all variables and interactions destined for the analysis model. The identification of these variables required special attention and was a computationally intensive exercise. S-CA, on the other hand, was found not to be affected by the missingness mechanism and, while the possibility exists and was illustrated, it is not necessary for any special inclusions – specifically in the form of interactions – to be made.

Compared to s-CA, MI methods are computationally complex and can be too intensive, depending on the number of variables included and the algorithm adopted. No problems with the number of variables included in the imputation model were experienced in this study, despite warnings that complications can arise when large numbers of variables are considered. It was found, however, that the inclusion of a full set of variables and the 10 identified interactions with s-CA, while not restrictive computationally, caused some crowding in the graphical display making its interpretation difficult. It was for this reason that the number of interactions in the comparative study was limited to two.

In many ways, the richness of inter-variable relationships from s-CA was superior to those from the MI procedure. With the application of s-CA all variables selected for inclusion in the study were retained in the analysis; whereas the development of the best model for analysis with MI and ordinal regression resulted in several variables being dropped altogether. Because of the sensitivity of model dependent analyses to the structure of the data, a specific problem was encountered in this study. With regard to the 'STOVE' variable, the incidence of 'gas stove' was only associated with the 'mild intermittent/ no asthma' level. This separation in the data caused problems during the application of ordinal regression and, consequently, the variable was dropped entirely from the analysis.

Because of model limitations in the application of ordinal regression following imputation, the asthma severity variable was reduced to three categories which caused some loss of information. This action was not necessary with s-CA and a finer distinction between the four levels of asthma severity was possible.

It was also found that, with s-CA, relationships between specific levels of asthma severity were identified and both within-variable category distinctions as well as between-variable category distinctions were possible. This enabled the identification of a specific set of variable categories associated with each asthma severity level to be made. With the MI application, on the other hand, it was only possible to identify associations between factors and one level of asthma severity relative to another level.

Statements regarding the significance of relationships were possible from analysis with MI when followed by ordinal regression and associations found could therefore be projected onto the broader population with a certain confidence. However, while strengths of associations between variables were suggested with s-CA, no significance, in the statistical sense, could be attached to these associations.

By selecting a set of core variables and interactions for application, a comparison between MI and s-CA was possible. Despite the differences outlined above regarding the requirements and application of these two methods, associations found between selected factors and asthma severity were consistent across methods. Furthermore, relationships found in this study between environmental, socio-economic, genetic and behavioural factors and asthma severity generally concur with those found in international studies. Higher levels of asthma severity were found to be associated with all forms of smoke exposure. Distinctions between the levels of asthma and the kinds of smoke exposure were also possible. From this study it can be concluded that the strongest association with moderate to severe asthma amongst this sample was found with smoke exposure in the form of industrial air pollution followed by smoke exposure in vehicles and then smoke exposure in the home. Smoke exposure *in utero* was found to have a closer association with mild persistent asthma.

Other factors, including gender, age and having had some form of neo-natal care were also shown to be associated with asthma severity. Internationally, results from different populations are not always consistent when considering the effect of socio-economic status on asthma prevalence and severity. Furthermore, there is growing debate on what constitutes a good measure for socio-economic status. Results from this study linked smaller families with worse asthma; in addition it was not the lowest income bracket that was shown to be associated with worse asthma, nor was it those who experienced violence in their neighbourhoods. It is hoped that these results which are specific to a region of South Africa which includes the highly industrialised South Durban basin can add to the knowledge of what factors affect not only the occurrence, but also the severity, of asthma in school-going children.

This study has clarified the diversity of two methods that can be used in the analysis of categorical data that suffer from missingness. Requirements, advantages and drawbacks have

been discussed for each method and, in an application of child asthma severity data, a consistency of results was confirmed.

On the basis of the findings of this study, both methods are equally reliable. It is up to the researcher to take into consideration limitations and requirements of the methods as well as the level of analysis required and the composition of the data and its missingness in order to make an informed choice of method that best serves his or her purpose.

While this study has illustrated the successful use of MI and s-CA in the management of missing data, some limitations and areas of future investigations have been identified.

Both MVNI and FCS were successfully applied to this set of categorical data with results being largely similar. Some minor discrepancies in significant levels relating to variables with more than two categories were, however, encountered. It is thought that these differences could be as a result of the imputation model used and therefore dependent on the assumed distribution of the imputation model. Further investigations would be useful to establish whether, in fact, the number of categories in a variable has a negative effect on the MVNI approach which traditionally assumes that variables follow a multinomial distribution.

The simulation study presented here was a simple one with a small selection of variables, only two of which were subjected to the pre-defined missingness scenarios. In addition, deletion ratios adopted were, in some cases, subjectively chosen. It would be a valuable exercise to expand the borders of this investigation by varying the number of variables imposed with missingness, as well as the ratios of these deletions. Whether deletions on inconsequential variables have the same effect as deletions on important variables would also be worthwhile exploring. While the small number of simulations applied in this study appeared to be adequate for the required accuracy of results, this could be increased as well to ascertain if an increase in simulations would radically change the outcomes.

# REFERENCES


ABAYOMI, K., GELMAN, A. & LEVY, M. 2008. Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 57**,** 273-291.

ALAIMO, K., OLSON, C. M., FRONGILLO JR, E. A. & BRIEFEL, R. R. 2001. Food insufficiency, family income, and health in US preschool and school-aged children. *American Journal of Public Health,* 91**,** 781-786.

ALLISON, P. D. 1987. Estimation of linear models with incomplete data. *Sociological Methodology,* 17**,** 71-103.

ALLISON, P. D. 1999. Logistic regression using the SAS system: theory and application. Cary, NC: SAS Institute Inc.

ALLISON, P. D. 2002. *Missing data,* Thousand Oaks, CA, Sage.

ALMQVIST, C., WORM, M. & LEYNAERT, B. 2007. Impact of gender on asthma in childhood and adolescence: a GA2LEN review. *Allergy,* 63**,** 47-57.

ALVES, G. D. C., SANTOS, D. N., FEITOSA, C. A. & BARRETO, M. L. 2012. Community violence and childhood asthma prevalence in peripheral neighborhoods in Salvador, Bahia State, Brazil. *Cadernos de Saúde Pública,* 28**,** 86-94.

ANDRIDGE, R. R. & LITTLE, R. J. 2010. A review of hot deck imputation for survey non-response. *International Statistical Review,* 78**,** 40-64.

APELBERG, B. J., AOKI, Y. & JAAKKOLA, J. J. 2001. Systematic review: Exposure to pets and risk of asthma and asthma-like symptoms. *Journal of Allergy and Clinical Immunology,* 107**,** 455-460.

ARBUCKLE, J. L., MARCOULIDES, G. A. & SCHUMACKER, R. E. 1996. Full information estimation in the presence of incomplete data. *In:* MARCOULIDES, G. A. & SCHUMACKER, R. E. (eds.) *Advanced Structural Equation Modeling: Issues and Techniques.* Mahwah, New Jersey: Lawrence Erlbaum, Inc.

ARBUCKLE, J. L. 2006. AMOS (version 7.0)[Computer software]. . Chicago, SPSS

ASHER, M. I., MONTEFORT, S., BJÖRKSTÉN, B., LAI, C. K., STRACHAN, D. P., WEILAND, S. K., WILLIAMS, H. & GROUP, I. P. T. S. 2006. Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys. *The Lancet,* 368**,** 733-743.

AZUR, M. J., STUART, E. A., FRANGAKIS, C. & LEAF, P. J. 2011. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research,* 20**,** 40-49.

BAE, H.-O., KIM, M. & HONG, S. M. 2008. Meal skipping children in low-income families and community practice implications. *Nutrition Research and Practice,* 2**,** 100-106.

BATES, M. N., CHANDYO, R. K., VALENTINER-BRANTH, P., POKHREL, A. K., MATHISEN, M., BASNET, S., SHRESTHA, P. S., STRAND, T. A. & SMITH, K. R. 2013. Acute lower respiratory infection in childhood and household fuel use in Bhaktapur, Nepal. *Environmental health perspectives,* 121**,** 637-642.

BECHER, R., HONGSLO, J. K., JANTUNEN, M. J. & DYBING, E. 1996. Environmental chemicals relevant for respiratory hypersensitivity: the indoor environment. *Toxicology Letters,* 86**,** 155-162.

BEDOLLA-BARAJAS, M., BARRERA-ZEPEDA, A. T., LÓPEZ-ZALDO, J. B. & MORALES-ROMERO, J. 2013. Asthma in Mexican school-age children is not associated with passive smoking or obesity. *Asia Pacific Allergy,* 3**,** 42-49.

BONNER, J. 1984. The epidemiology and natural history of asthma. *Clinics in Chest Medicine,* 5**,** 557-565.

BRANT, R. 1990. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics***,** 1171-1178.

BROOKS, A.-M., BYRD, R. S., WEITZMAN, M., AUINGER, P. & MCBRIDE, J. T. 2001. Impact of low birth weight on early childhood asthma in the United States. *Archives of Pediatrics & Adolescent Medicine,* 155**,** 401-406.

BURTON, A., ALTMAN, D. G., ROYSTON, P. & HOLDER, R. L. 2006. The design of simulation studies in medical statistics. *Statistics in Medicine,* 25**,** 4279-4292.

CARLSEN, K. C. L., ROLL, S., CARLSEN, K.-H., MOWINCKEL, P., WIJGA, A. H., BRUNEKREEF, B., TORRENT, M., ROBERTS, G., ARSHAD, S. H. & KULL, I. 2012. Does pet ownership in infancy lead to asthma or allergy at school age? Pooled analysis of individual participant data from 11 European birth cohorts. *PloS ONE,* 7**,** e43214.

CESARONI, G., FARCHI, S., DAVOLI, M., FORASTIERE, F. & PERUCCI, C. 2003. Individual and area-based indicators of socioeconomic status and childhood asthma. *European Respiratory Journal,* 22**,** 619-624.

CHAROENCA, N., KUNGSKULNITI, N., TIPAYAMONGKHOLGUL, M., SUJIRARAT, D., LOHCHINDARAT, S., MOCK, J. & HAMANN, S. L. 2013. Determining the burden of secondhand smoke exposure on the respiratory health of Thai children. *Tobacco Induced Diseases,* 11**,** 7-12.

CHARPIN, D., KLEISBAUER, J., FONDARAI, J., GRALAND, B., VIALA, A. & GOUEZO, F. 1988. Respiratory symptoms and air pollution changes in children: the Gardanne coal-basin study. *Archives of Environmental Health: An International Journal,* 43**,** 22-27.

CHEN, E., CHIM, L. S., STRUNK, R. C. & MILLER, G. E. 2007. The role of the social environment in children and adolescents with asthma. *American Journal of Respiratory and Critical Care Medicine,* 176**,** 644-649.

CHOI, W.-J., UM, I.-Y., HONG, S., YUM, H. Y., KIM, H. & KWON, H. 2012. Association between household income and asthma symptoms among elementary school children in Seoul. *Environmental Health and Toxicology,* 27**,** e2012020.

COHEN, J., COHEN, P., WEST, S. G. & AIKEN, L. S. 2003. *Applied multiple regression/correlation analysis for the behavioral sciences,* Mahwah, New Jersey, Lawrence Erlbaum Associates, Inc.

COLLINS, L. M., SCHAFER, J. L. & KAM, C.-M. 2001. A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods,* 6**,** 330-351.

CORBO, G. M., FORASTIERE, F., DE SARIO, M., BRUNETTI, L., BONCI, E., BUGIANI, M., CHELLINI, E., LA GRUTTA, S., MIGLIORE, E. & PISTELLI, R. 2008. Wheeze and asthma in children: associations with body mass index, sports, television viewing, and diet. *Epidemiology,* 19**,** 747-755.

DAVIS, J. B. & BULPITT, C. J. 1981. Atopy and wheeze in children according to parental atopy and family size. *Thorax,* 36**,** 185-189.

DEGER, L., PLANTE, C., JACQUES, L., GOUDREAU, S., PERRON, S., HICKS, J., KOSATSKY, T. & SMARGIASSI, A. 2012. Active and uncontrolled asthma among children exposed to air stack emissions of sulphur dioxide from petroleum refineries in Montreal, Quebec: a cross-sectional study. *Canadian Respiratory Journal: Journal of the Canadian Thoracic Society,* 19**,** 97-102.

DEMIRTAS, H. & SCHAFER, J. L. 2003. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine,* 22**,** 2553-2575.

DEMIRTAS, H. 2005. Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine,* 24**,** 2345-2363.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1-38.

DESAI, M., ESSERMAN, D. A., GAMMON, M. D. & TERRY, M. B. 2011. The use of complete-case and multiple imputation-based analyses in molecular epidemiology studies that assess interaction effects. *Epidemiologic Perspectives & Innovations,* 8**,** 1-17.

DIFRANZA, J. R., ALIGNE, C. A. & WEITZMAN, M. 2004. Prenatal and postnatal environmental tobacco smoke exposure and children's health. *Pediatrics,* 113**,** 1007-1015.

DIK, N., TATE, R. B., MANFREDA, J. & ANTHONISEN, N. R. 2004. Risk of physician-diagnosed asthma in the first 6 years of life. *CHEST Journal,* 126**,** 1147-1153.

DONDERS, A. R. T., VAN DER HEIJDEN, G. J., STIJNEN, T. & MOONS, K. G. 2006. Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology,* 59**,** 1087-1091.

EEKHOUT, I., DE BOER, R. M., TWISK, J. W., DE VET, H. C. & HEYMANS, M. W. 2012. Missing Data: A Systematic Review of How They Are Reported and Handled. *Epidemiology,* 23**,** 729-732.

EHRLICH, R., KATTAN, M., GODBOLD, J., SALTZBERG, D. S., GRIMM, K. T., LANDRIGAN, P. & LILIENFELD, D. 1992. Childhood asthma and passive smoking. *American Review of Respiratory Diseases,* 145**,** 594-599.

ENDERS, C. K. 2001. A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling,* 8**,** 128-141.

ENDERS, C. K. & BANDALOS, D. L. 2001. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling,* 8**,** 430-457.

ERNST, P., DEMISSIE, K., JOSEPH, L., LOCHER, U. & BECKLAKE, M. R. 1995. Socioeconomic status and indicators of asthma in children. *American Journal of Respiratory and Critical Care Medicine,* 152**,** 570-575.

FARIS, P. D., GHALI, W. A., BRANT, R., NORRIS, C. M., GALBRAITH, P. D., KNUDTSON, M. L. & INVESTIGATORS, A. 2002. Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. *Journal of Clinical Epidemiology,* 55**,** 184-191.

FINCH, W. H. 2010. Imputation methods for missing categorical questionnaire data: A comparison of approaches. *Journal of Data Science,* 8**,** 361-378.

FISHER, R. A. 1940. The precision of discriminant functions. *Annals of Eugenics,* 10**,** 422-429.

FORD, B. M. 1983. An Overview of Hot-Deck Procedures. *In:* MADOW, W., OLKIN, I. & RUBIN, D. (eds.) *Incomplete Data in Sample Surveys, vol 2, Theory and Bibliographies.* New York: Academic Press.

FORNO, E. & CELEDÓN, J. C. 2009. Asthma and ethnic minorities: socioeconomic status and beyond. *Current Opinion in Allergy and Clinical Immunology,* 9**,** 154-160.

GARSON, G. D. 2008. Ordinal regression. Statnotes: Topics in multivariate analysis.

GENT, J. F., BELANGER, K., TRICHE, E. W., BRACKEN, M. B., BECKETT, W. S. & LEADERER, B. P. 2009. Association of pediatric asthma severity with exposure to common household dust allergens. *Environmental Research,* 109**,** 768-774.

GOH, D., CHEW, F., QUEK, S. & LEE, B. 1996. Prevalence and severity of asthma, rhinitis, and eczema in Singapore schoolchildren. *Archives of Disease in Childhood,* 74**,** 131-135.

GOLD, D. R. & WRIGHT, R. 2005. Population disparities in asthma. *Annual Review of Public Health,* 26**,** 89-113.

GOLDBERG, S., ISRAELI, E., SCHWARTZ, S., SHOCHAT, T., IZBICKI, G., TOKER-MAIMON, O., KLEMENT, E. & PICARD, E. 2007. Asthma prevalence, family size, and birth order. *CHEST Journal,* 131**,** 1747-1752.

GOODWIN, R. D. & COWLES, R. A. 2008. Household smoking and childhood asthma in the United States: a state-level analysis. *Journal of Asthma,* 45**,** 607-610.

GRAHAM, J. W. & DONALDSON, S. I. 1993. Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology,* 78**,** 119-128.

GRAHAM, J. W., HOFER, S. M., DONALDSON, S. I., MACKINNON, D. P. & SCHAFER, J. L. 1997. Analysis with missing data in prevention research. *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research,* 1**,** 325-366.

GRAHAM, J. W. & SCHAFER, J. L. 1999. On the performance of multiple imputation for multivariate data with small sample size. *Statistical Strategies for Small Sample Research,* 50**,** 1-27.

GRAHAM, J. W., CUMSILLE, P. E. & ELEK-FISK, E. 2003. Methods for handling missing data. *In:* WEINER, I. B. (ed.) *Handbook of Psychology.* Wiley online.

GRAHAM, J. W., OLCHOWSKI, A. E. & GILREATH, T. D. 2007. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science,* 8**,** 206-213.

GRAHAM, J. W. 2009. Missing data analysis: Making it work in the real world. *Annual Review of Psychology,* 60**,** 549-576.

GRAHAM, J. W. 2012. *Missing data: Analysis and design,* New York, Springer.

GREENACRE, M. J. 1978. Some Objective Methods of Graphical Display of a Data Matrix: Special Report. Pretoria: Department of Statistics and Operations Research, University of South Africa.

GREENACRE, M. J. 1984. *Theory And Applications Of Correspondence Analysis,* London, Academic Press.

GREENACRE, M. J. 1992. Correspondence analysis in medical research. *Statistical Methods in Medical Research,* 1**,** 97-117.

GREENACRE, M. J. & BLASIUS, J. 2006. Correspondence Analysis and Related Methods in Practice. *In:* GREENACRE, M. J. & BLASIUS, J. (eds.) *Multiple Correspondence Analysis and Related Methods.* Boca Raton: Chapman & Hall/CRC.

GREENACRE, M. J. & PARDO, R. 2006a. Multiple Correspondence Analysis of Subsets of Response Categories. *In:* GREENACRE, M. & BLASIUS, J. (eds.) *Multiple Correspondence Analysis and Related Methods.* Boca Raton: Chapman & Hall/CRC.

GREENACRE, M. J. & PARDO, R. 2006b. Subset correspondence analysis visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods & Research,* 35**,** 193-218.

GREENLAND, S. & FINKLE, W. D. 1995. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology,* 142**,** 1255-1264.

GUTTMAN, L. 1941. The quantification of a class of attributes: A theory and method of scale construction. *The Prediction of Personal Adjustment***,** 319-348.

HACK, M., TAYLOR, H. G., DROTAR, D., SCHLUCHTER, M., CARTAR, L., ANDREIAS, L., WILSON-COSTELLO, D. & KLEIN, N. 2005. Chronic conditions, functional limitations, and special health care needs of school-aged children born with extremely low-birth-weight in the 1990s. *The Journal of the American Medical Association,* 294**,** 318-325.

HANCOX, R. J., MILNE, B. J., TAYLOR, D., GREENE, J. M., COWAN, J. O., FLANNERY, E. M., HERBISON, G., MCLACHLAN, C. R., POULTON, R. & SEARS, M. R. 2004. Relationship between socioeconomic status and asthma: a longitudinal cohort study. *Thorax,* 59**,** 376-380.

HARDT, J. & GÖRGEN, K. Multiple imputation using ICE: A simulation study on a binary response. German Stata Users' Group Meetings 2008. Stata Users Group.

HARDT, J., HERKE, M. & LEONHART, R. 2012. Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC Medical Research Methodology,* 12**,** 1-13.

HARJU, T., KEISTINEN, T., TUUPONEN, T. & KIVELÄ, S. L. 1996. Hospital admissions of asthmatics bv age and sex. *Allergy,* 51**,** 693-696.

HAYASHI, C. 1950. On the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics,* 2**,** 35-47.

HE, Y. 2010. Missing Data Analysis Using Multiple Imputation Getting to the Heart of the Matter. *Circulation: Cardiovascular Quality and Outcomes,* 3**,** 98-105.

HECKMAN, J. J. 1979. Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society***,** 153-161.

HENDRY, G., NORTH, D., ZEWOTIR, T. & NAIDOO, R. 2014a. The application of subset correspondence analysis to address the problem of missing data in a study on asthma severity in childhood. *Statistics in Medicine,* 33**,** 3882-3893.

HENDRY, G. M., NAIDOO, R. N., ZEWOTIR, T., NORTH, D. & MENTZ, G. 2014b. Model development including interactions with multiple imputed data. *BMC Medical Research Methodology,* 14**,** 1-11.

HENDRY, G. M., ZEWOTIR, T., NAIDOO, R. N. & NORTH, D. 2015. The Effect of the Mechanism and Amount of Missingness on Subset Correspondence Analysis. *Correspondence in Statistics,* Under review.

HENRY, R. L., ABRAMSON, R., ADLER, J. A., WLODARCYZK, J. & HENSLEY, M. J. 1991. Asthma in the vicinity of power stations: I. A prevalence study. *Pediatric Pulmonology,* 11**,** 127-133.

HIRSCHFELD, H. O. A connection between correlation and contingency. Mathematical Proceedings of the Cambridge Philosophical Society, 1935. Cambridge Univ Press, 520-524.

HORST, P. 1935. Measuring complex attitudes. *The Journal of Social Psychology,* 6**,** 369-374.

HOWELL, D. C. 2007. The treatment of missing data. *In:* OUTHWAITE, W. & TURNER, S. (eds.) *Handbook of Social Science Methodology.* London: Sage.

HSU, H.-H. L., CHIU, Y.-H. M., COULL, B. A., KLOOG, I., SCHWARTZ, J., LEE, A., WRIGHT, R. O. & WRIGHT, R. J. 2015. Prenatal Particulate Air Pollution and Asthma Onset in Urban Children: Identifying Sensitive Windows and Sex Differences. *American Journal of Respiratory and Critical Care Medicine,* In press.

JARVIS, D., CHINN, S., LUCZYNSKA, C. & BURNEY, P. 1997. The association of family size with atopy and atopic disease. *Clinical & Experimental Allergy,* 27**,** 240-245.

JEFFREY, J., STERNFELD, I. & TAGER, I. 2006. The association between childhood asthma and community violence, Los Angeles County, 2000. *Public Health Reports,* 121**,** 720-728.

JONES, A. P. 1998. Asthma and domestic air quality. *Social Science & Medicine,* 47**,** 755-764.

JONES, M. P. 1996. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association,* 91**,** 222-230.

KAPLAN, B. A. & MASCIE-TAYLOR, C. 1985. Biosocial factors in the epidemiology of childhood asthma in a British national sample. *Journal of Epidemiology and Community Health,* 39**,** 152-156.

KARAHALIOS, A., BAGLIETTO, L., CARLIN, J. B., ENGLISH, D. R. & SIMPSON, J. A. 2012. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC Medical Research Methodology,* 12**,** 1-10.

KARMAUS, W. & BOTEZAN, C. 2002. Does a higher number of siblings protect against the development of allergy and asthma? A review. *Journal of Epidemiology and Community Health,* 56**,** 209-217.

KLEBANOFF, M. A. & COLE, S. R. 2008. Use of multiple imputation in the epidemiologic literature. *American Journam of Epidemiology,* 168**,** 355-357.

KOZYRSKYJ, A. L., KENDALL, G. E., JACOBY, P., SLY, P. D. & ZUBRICK, S. R. 2010. Association between socioeconomic status and the development of asthma: analyses of income trajectories. *American Journal of Public Health,* 100**,** 540-546.

KWON, H. L., ORTIZ, B., SWANER, R., SHOEMAKER, K., JEAN-LOUIS, B., NORTHRIDGE, M. E., VAUGHAN, R. D., MARX, T., GOODMAN, A. & BORRELL, L. N. 2006. Childhood asthma and extreme values of body mass index: the Harlem Children's Zone Asthma Initiative. *Journal of Urban Health,* 83**,** 421-433.

LAI, C., DOUGLASS, C., HO, S., LAU, J., WONG, G. & LEUNG, R. 1996. Asthma epidemiology in the Far East. *Clinical & Experimental Allergy,* 26**,** 5-12.

LEE, K. J. & CARLIN, J. B. 2010. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology,* 171**,** 624-632.

LITONJUA, A. A., CAREY, V. J., WEISS, S. T. & GOLD, D. R. 1999. Race, socioeconomic factors, and area of residence are associated with asthma prevalence. *Pediatric Pulmonology,* 28**,** 394-401.

LITTLE, R. J. & RUBIN, D. B. 1987. *Statistical Analysis With Missing Data,* New York, Wiley.

LITTLE, R. J. 1995. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association,* 90**,** 1112-1121.

LITTLE, R. J. & SCHENKER, N. 1995. Missing data. *Handbook of Statistical Modeling for the Social and Behavioral Sciences.* New York: Springer.

LU, F. L., HSIEH, C.-J., CAFFREY, J. L., LIN, M.-H., LIN, Y.-S., LIN, C.-C., TSAI, M.-S., HO, W.-C., CHEN, P.-C. & SUNG, F.-C. 2012. Body mass index may modify asthma prevalence among low-birth-weight children. *American Journal of Epidemiology,* 176**,** 32-42.

LYNCH, S. 2003. Missing data (Soc 504). Princeton University Sociology 504 Class Notes. Princeton University.

MAANTAY, J. 2007. Asthma and air pollution in the Bronx: methodological and data considerations in using GIS for environmental justice and health research. *Health & Place,* 13**,** 32-56.

MAI, X. M., GÄDDLIN, P. O., NILSSON, L., FINNSTRÖM, O., BJÖRKSTÉN, B., JENMALM, M. C. & LEIJON, I. 2003. Asthma, lung function and allergy in 12-year-old children with very low birth weight: A prospective study. *Pediatric Allergy and Immunology,* 14**,** 184-192.

MARSHALL, A., ALTMAN, D. G., ROYSTON, P. & HOLDER, R. L. 2010. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Medical Research Methodology,* 10**,** 1-16.

MATOOANE, L. & DIAB, R. 2001. Air pollution carrying capacity in the South Durban Industrial Basin: research in action. *South African Journal of Science,* 97**,** 450-453.

MATRICARDI, P. M., FRANZINELLI, F., FRANCO, A., CAPRIO, G., MURRU, F., CIOFFI, D., FERRIGNOC, L., PALERMOA, A., CICCARELLI, N. & ROSMINI, F. 1998. Sibship size, birth order, and atopy in 11,371 Italian young men. *Journal of Allergy and Clinical Immunology,* 101**,** 439-444.

MCKEEVER, T., LEWIS, S., SMITH, C., COLLINS, J., HEATLIE, H., FRISCHER, M. & HUBBARD, R. 2001. Siblings, multiple births, and the incidence of allergic disease: a birth cohort study using the West Midlands general practice research database. *Thorax,* 56**,** 758-762.

MCMANUS, B. M., ROBERT, S., ALBANESE, A., SADEK-BADAWI, M. & PALTA, M. 2012. Racial disparities in health-related quality of life in a cohort of very-low-birth-weight 2-and 3-year-olds with and without asthma. *Journal of Epidemiology and Community Health,* 66**,** 579-585.

MOLNAR, F. J., HUTTON, B. & FERGUSSON, D. 2008. Does analysis using 'last observation carried forward' introduce bias in dementia research? *Canadian Medical Association Journal,* 179**,** 751-753.

MOORE, J. C., STINSON, L. L. & WELNIAK, E. J. 2000. Income measurement error in surveys: a review. *Journal of Official Statistics - Stockolm,* 16**,** 331-362.

MUTHÉN, B., KAPLAN, D. & HOLLIS, M. 1987. On structural equation modeling with data that are not missing completely at random. *Psychometrika,* 52**,** 431-462.

NAEPP 1991. National Asthma Education Prevention Program. Expert Panel Report: Guidelines for the Diagnosis and Management of Asthma. Bethesda, MD: National Institutes of Health.

NAFSTAD, P., BRUNEKREEF, B., SKRONDAL, A. & NYSTAD, W. 2005. Early respiratory infections, asthma, and allergy: 10-year follow-up of the Oslo Birth Cohort. *Pediatrics,* 116**,** 255-262.

NAIDOO, R. N., ROBINS, T. G., BATTERMAN, S., MENTZ, G. & JACK, C. 2013. Ambient pollution and respiratory outcomes among schoolchildren in Durban, South Africa. *South African Journal of Child Health,* 7**,** 127-134.

NEIDELL, M. J. 2004. Air pollution, health, and socio-economic status: the effect of outdoor air quality on childhood asthma. *Journal of Health Economics,* 23**,** 1209-1236.

NOVO, A. & SCHAFER, J. 2010. norm: Analysis of multivariate normal datasets with missing values.

O'CONNELL, A. A. 2006. *Logistic regression models for ordinal response variables,* Thousand Oaks, Sage.

PECKHAM, C. & BUTLER, N. 1978. A national study of asthma in childhood. *Journal of Epidemiology and Community Health,* 32**,** 79-85.

PEDEN, D. 2003. Air pollution: indoor and outdoor. *In:* ADKINSON NF JR, Y. J., BUSSE WW, BOCHNER BS, HOLGATE SK, SIMONS FE (ed.) *Middleton's Allergy: Principles and Practice.* Philadelphia: Mosby.

PEDEN, D. B. 2005. The epidemiology and genetics of asthma risk associated with air pollution. *Journal of Allergy and Clinical Immunology,* 115**,** 213-219.

PERSKY, V. W., SLEZAK, J., CONTRERAS, A., BECKER, L., HERNANDEZ, E., RAMAKRISHNAN, V. & PIORKOWSKI, J. 1998. Relationships of race and socioeconomic status with prevalence, severity, and symptoms of asthma in Chicago school children. *Annals of Allergy, Asthma & Immunology,* 81**,** 266-271.

PEYRE, H., LEPLÈGE, A. & COSTE, J. 2011. Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of Life Research,* 20**,** 287-300.

PITTMAN, T. P., NYKIFORUK, C. I., MIGNONE, J., MANDHANE, P. J., BECKER, A. B. & KOZYRSKYJ, A. L. 2012. The association between community stressors and asthma prevalence of school children in Winnipeg, Canada. *International Journal of Environmental Research and Public Health,* 9**,** 579-595.

PLEIS, J. R. & COHEN, R. A. 2007. Impact of income bracketing on poverty measures used in the National Health Interview Survey's Early Release Program: Preliminary data from the 2007 NHIS. Hyattsville, MD: National Center for Health Statistics. December 2007. *Hyattsville, MD: National Center for Health Statistics*.

POYSER, M., NELSON, H., EHRLICH, R., BATEMAN, E., PARNELL, S., PUTERMAN, A. & WEINBERG, E. 2002. Socioeconomic deprivation and asthma prevalence and severity in young adolescents. *European Respiratory Journal,* 19**,** 892-898.

RAGHUNATHAN, T. E., SOLENBERGER, P. W. & VAN HOEWYK, J. 2002. IVEware: Imputation and variance estimation software. *Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan.* Ann Arbor, MI.

RAUH, V. A., LANDRIGAN, P. J. & CLAUDIO, L. 2008. Housing and Health. *Annals of the New York Academy of Sciences,* 1136**,** 276-288.

RAYENS, M. K., BURKHART, P. V., ZHANG, M., LEE, S., MOSER, D. K., MANNINO, D. & HAHN, E. J. 2008. Reduction in asthma-related emergency department visits after implementation of a smoke-free law. *Journal of Allergy and Clinical Immunology,* 122**,** 537-541.

RDEVELOPMENT_CORE_TEAM 2006. *R: A language and environment for statistical computing.* Vienna, Austria.

RICHARDSON, M. & KUDER, G. 1933. Making a rating scale that measures. *Personnel* 12**,** 36-40.

RITTER, F. E., SCHOELLES, M. J., QUIGLEY, K. S. & KLEIN, L. C. 2011. Determining the number of simulation runs: Treating simulations as theories by not sampling their behavior. *Human-in-the-Loop Simulations.* London: Springer.

ROBERTS, W. & EHRLICH, R. 2009. Meteorologically estimated exposure but not distance predicts asthma symptoms in schoolchildren in the environs of a petrochemical refinery: a cross-sectional study. *Environmental Health,* 8**,** 1-10.

ROSENSTREICH, D. L., EGGLESTON, P., KATTAN, M., BAKER, D., SLAVIN, R. G., GERGEN, P., MITCHELL, H., MCNIFF-MORTIMER, K., LYNN, H. & OWNBY, D. 1997. The role of cockroach allergy and exposure to cockroach allergen in causing morbidity among inner-city children with asthma. *New England Journal of Medicine,* 336**,** 1356-1363.

RUBIN, D. B. 1976. Inference and missing data. *Biometrika,* 63**,** 581-592.

RUBIN, D. B. 1996. Multiple imputation after 18+ years. *Journal of the American Statistical Association,* 91**,** 473-489.

RUBIN, D. B. 2004a. The design of a general and flexible system for handling nonresponse in sample surveys. *The American Statistician,* 58**,** 298-302.

RUBIN, D. B. 2004b. *Multiple imputation for nonresponse in surveys,* New York, Wiley

SCHAFER, J. L. 1997. *Analysis of Incomplete Multivariate Data,* New York, Chapman & Hall.

SCHAFER, J. L. & OLSEN, M. K. 1998. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research,* 33**,** 545-571.

SCHAFER, J. L. 1999. NORM: Multiple imputation of incomplete multivariate data under a normal model [Computer software]. University Park: Pennsylvania State University, Department of Statistics.

SCHAFER, J. L. & GRAHAM, J. W. 2002. Missing data: our view of the state of the art. *Psychological Methods,* 7**,** 147.

SCHEFFER, J. 2002. Dealing with missing data. *Research Letters in the Information and Mathematical Sciences,* 3**,** 153-160.

SCHEUREN, F. 2005. Multiple imputation: How it began and continues. *The American Statistician,* 59**,** 315-319.

SENDZIK, T., FONG, G. T., TRAVERS, M. J. & HYLAND, A. 2009. An experimental investigation of tobacco smoke pollution in cars. *Nicotine & Tobacco Research,* 11**,** 627-634.

SHANKARDASS, K., MCCONNELL, R. S., MILAM, J., BERHANE, K., TATALOVICH, Z., WILSON, J. P. & JERRETT, M. 2007. The association between contextual socioeconomic factors and prevalent asthma in a cohort of Southern California school children. *Social Science & Medicine,* 65**,** 1792-1806.

SHAVERS, V. L. 2007. Measurement of socioeconomic status in health disparities research. *Journal of the National Medical Association,* 99**,** 1013-1023.

SHRIVE, F. M., STUART, H., QUAN, H. & GHALI, W. A. 2006. Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. *BMC Medical Research Methodology,* 6**,** 57.

SLY, P. D., DEVERELL, M., KUSEL, M. M. & HOLT, P. G. 2007. Exposure to environmental tobacco smoke in cars increases the risk of persistent wheeze in adolescents. *Medical Journal of Australia,* 186**,** 322-322.

SPSS INC. Build Better Models When You Fill in the Blanks. Available from: http://www.spss.com/media/collateral/statistics/missing-values.pdfs (20 April 2014).

STRACHAN, D. P. 1989. Hay fever, hygiene, and household size. *British Medical Journal,* 299**,** 1259 - 1260.

STUART, E. A., AZUR, M., FRANGAKIS, C. & LEAF, P. 2009. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *American Jornal of Epidemiology,* 169**,** 1133-1139.

SUBRAMANIAN, S. & KENNEDY, M. H. 2009. Perception of neighborhood safety and reported childhood lifetime asthma in the United States (US): a study based on a national survey. *PloS ONE,* 4**,** e6091.

TAKKOUCHE, B., GONZÁLEZ-BARCALA, F. J., ETMINAN, M. & FITZGERALD, M. 2008. Exposure to furry pets and the risk of asthma and allergic rhinitis: a meta-analysis. *Allergy,* 63**,** 857-864.

TANNER, M. A. & WONG, W. H. 1987. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association,* 82**,** 528-540.

TORRES-LACOMBA, A. 2006. Correspondence analysis and categorical conjoint measurement. *In:* GREENACRE, M. J. & BLASIUS, J. (eds.) *Multiple Correspondence Analysis and Related Methods.* Boca Raton: Chapman & Hall/CRC.

VACH, W. & BLETTNER, M. 1991. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American Journal of Epidemiology,* 134**,** 895-907.

VAN BUUREN, S. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research,* 16**,** 219-242.

VAN BUUREN, S. & GROOTHUIS-OUDSHOORN, K. 2011. MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software,* 45**,** 1-68.

VENABLES, K. M. & CHAN-YEUNG, M. 1997. Occupational asthma. *The Lancet,* 349**,** 1465-1469.

VERGOUWE, Y., ROYSTON, P., MOONS, K. G. & ALTMAN, D. G. 2010. Development and validation of a prediction model with missing predictor data: a practical approach. *Journal of Clinical Epidemiology,* 63**,** 205-214.

VON HIPPEL, P. T. 2009. How to impute interactions, squares, and other transformed variables. *Sociological Methodology,* 39**,** 265-291.

VON MUTIUS, E. 2000. The environmental predictors of allergic disease. *Journal of Allergy and Clinical Immunology,* 105**,** 9-19.

WHITE, I. R., ROYSTON, P. & WOOD, A. M. 2011. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine,* 30**,** 377-399.

WILLIAMSON, I., MARTIN, C., MCGILL, G., MONIE, R. & FENNERTY, A. 1997. Damp housing and asthma: a case-control study. *Thorax,* 52**,** 229-234.

WOOD, A. M., WHITE, I. R. & THOMPSON, S. G. 2004. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials,* 1**,** 368-376.

WOOD, A. M., WHITE, I. R. & ROYSTON, P. 2008. How should variable selection be performed with multiply imputed data? *Statistics in Medicine,* 27**,** 3227-3246.

WRIGHT, R. J., MITCHELL, H., VISNESS, C. M., COHEN, S., STOUT, J., EVANS, R. & GOLD, D. R. 2004. Community violence and asthma morbidity: the Inner-City Asthma Study. *American Journal of Public Health,* 94**,** 625-632.

WRIGHT, R. J. 2008. Stress and childhood asthma risk: overlapping evidence from animal studies and epidemiologic research. *Allergy, Asthma and Clinical Immunology,* 4**,** 29-36.

# APPENDICES: Published and submitted papers

Research Article

Statistics in Medicine

Received 10 December 2012, Accepted 6 April 2014 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6189

# The application of subset correspondence analysis to address the problem of missing data in a study on asthma severity in childhood

**G. Hendry,[a*†] D. North,[a] T. Zewotir[a] and R. N. Naidoo[b]**

**Non-response in cross-sectional data is not uncommon and requires careful handling during the analysis stage so as not to bias results. In this paper, we illustrate how subset correspondence analysis can be applied in order to manage the non-response while at the same time retaining all observed data. This variant of correspondence analysis was applied to a set of epidemiological data in which relationships between numerous environmental, genetic, behavioural and socio-economic factors and their association with asthma severity in children were explored. The application of subset correspondence analysis revealed interesting associations between the measured variables that otherwise may not have been exposed. Many of the associations found confirm established theories found in literature regarding factors that exacerbate childhood asthma. Moderate to severe asthma was found to be associated with needing neonatal care, male children, 8- to 9-year olds, exposure to tobacco smoke in vehicles and living in areas that suffer from extreme air pollution. Associations were found between mild persistent asthma and low birthweight, and being exposed to smoke in the home and living in a home with up to four people. The classification of probable asthma was associated with a group of variables that indicate low socio-economic status.** **Copyright © 2014 John Wiley & Sons, Ltd.**

**Keywords:** asthma severity; categorical data analysis; missing data; subset correspondence analysis; supplementary variables; total inertia

## 1. Introduction

Missing data is an ongoing challenge for many researchers and presents a particular problem in community-based epidemiological studies. It is evident that missing data is still frequently being handled by ad hoc methods, such as complete case analysis [1]. If the missing data are categorical, an extra 'missing' category is sometimes added for each incomplete variable. These methods of dealing with missing data may, however, result in biased estimates and are thus not recommended [2, 3]. Multiple imputation, a tool that is becoming more popular for dealing with missing data, is often used in conjunction with some regression procedures to analyse multivariate data that suffer from missingness. These aforementioned methods of handling missing data are all sensitive to the missingness mechanism present in the data. Furthermore, their use is often restricted by complexities of models and distributional requirements. We believe that a more favourable approach is the application of correspondence analysis (CA), and its variants, which are commonly used in the analysis of multivariate categorical data. These methods do not assume a model and are therefore not restricted by distributional requirements nor are they sensitive to the missingness mechanism present in the data. Specifically, this paper will focus on subset CA (s-CA) as an exploratory tool to deal with missingness when the data are categorical. This procedure thus takes an alternative approach to analyse data (that have missing values) than the conventional methodology classically favoured by epidemiologists.

[a] *School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa*
[b] *Discipline of Occupational and Environmental Health, University of KwaZulu-Natal, Durban, South Africa*
**Correspondence to: G. Hendry, School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa.*
[†] *E-mail: hendryfam@telkomsa.net*

Copyright © 2014 John Wiley & Sons, Ltd. 

The brainchild of Benzécri, CA, originated in France in the early 1960s and is gaining popularity as an exploratory tool for analysing multivariate categorical data [4]. CA is primarily a graphical technique used to explore the relationships between variables. When the number of variables in a study is large, CA can be used as a tool to select important variables to consider for further analysis. Unlike the more classical regression-based methods for studying inter-variable relationships that hypothesise a model and fit the data to a model, the extended family of methods under CA do not hypothesise a model. Instead, the data are decomposed in order to study their 'structure' [5]. Points (rows and columns of a data matrix), represented as clouds in multi-dimensional space, are optimally displayed in a lower dimensional subspace, which is easier to interpret because of the lower dimensionality.

Although it is usual to apply CA to the full set of data, there are times when the analysis of a subset of the data may be more appropriate or desirable: for instance, when one is only interested in the analysis of agreement scores on a Likert-scale-type questionnaire. Another setting for opting for the analysis of a reduced set of the data is when the data set has a large number of variables, often further broken down into categories. In this case, the interpretation of the plots can become complicated because of over-crowding. All the variables/categories load to some extent on all dimensions, and it is usually not possible to obtain more than a broad overview of, often expected, relationships [6]. More information and insight could thus be gained into associations of variables if smaller groups were analysed individually.

This has been made possible by the development of s-CA , a variant of CA [3]. In s-CA, as the name suggests, a subset of the data matrix is selected for analysis. CA is then applied with the important modification that the marginal frequencies of the full matrix are retained in the analysis of the subset.

While the application of s-CA had been documented illustrating the analysis of the subset of observed responses in a study with Likert scale data [7], there is little evidence in the literature of its application to studies with missing data.

We will illustrate s-CA on a set of epidemiological data with a large number of variables in which missingness is present, from a study of asthma severity in children in Durban, South Africa. The non-response for each variable was categorised separately, and the subset of observed categories was analysed. This method offers a way of dealing with missing categorical data while, at the same time, retaining all records, complete and incomplete. We believe this complementary approach is a better choice for the analysis of categorical data that suffer from missingness, as it is simple to apply and circumvents the model approximations and missingness mechanism dilemma.

## 2. Theoretical concepts

### *2.1. Correspondence analysis of a subset of the data*

Correspondence analysis is an exploratory multivariate technique applied to any matrix of non-negative numbers in order to identify associations present in the data. In CA, the rows and columns of the matrix are represented by two separate clouds of points in multi-dimensional space. CA finds respective subspaces of low dimension that optimally contain these clouds of points. The principal axes are chosen such that the inertia of the clouds of points is maximised. The inertia of these clouds can be considered as a measure of dispersion or spread of the points taking into account both distance and attributed weights, called masses. CA thus provides a visual interpretation of the relative positions of both clouds in a common subspace of low dimension. Interpretation of the axes can be achieved by examining the decomposition of the inertia of each cloud of points along the principal axes and amongst the points themselves [5]. By studying the contributions that the points make to the principal axes and the contributions that the axes make to the inertia of the points, those points that are well defined in a plane can be identified. Using these points, it is usually possible to assign 'meanings' to the principal axes. Graphically, if the angle between this point vector and the axis is small, then the point is highly correlated with the principal axis. The distance between two points (either two row points or two column points) is said to be a 'weak' approximation of the chi-square distance between the vectors of relative frequencies of the points [8]. One can get an idea of how close two points are by examining the angle the point vectors make with each other. The smaller the angle, the closer they are related. The interpretation of the graphical display is primarily carried out on the basis of where a point, or group of points, is positioned relative to the axes in the plane.

The variables used in the calculation of the subspace are called active variables. It is possible to examine the position of additional variables, called supplementary variables, relative to this space. These variables play no part in the determination of the principal axes and the optimal subspace but

Copyright © 2014 John Wiley & Sons, Ltd.

are projected onto an existing subspace. Relationships between these variables, both active and supplementary, and the principal axes can be explored [4, 5]. In practice, the associations of the active variables are displayed, and then, the supplementary variables are related a posteriori to these associations [6, 9].

In the same way that CA is applied to a full set of data, s-CA is applied to a subset of the data. An appealing feature of s-CA is that, as the full data matrix, $\boldsymbol{N}$, can be partitioned into a number of separate non-overlapping and all-inclusive matrices, so is the inertia of the full matrix equal to the sum of the inertias of the separate matrices [7].

So, if $\boldsymbol{N} = [\boldsymbol{N}_1 : \boldsymbol{N}_2 : \boldsymbol{N}_3]$ , it follows that the inertia of $\boldsymbol{N}$, In($\boldsymbol{N}$), follows the rule

$$\text{In}(\boldsymbol{N}) = \text{In}(\boldsymbol{N}_1) + \text{In}(\boldsymbol{N}_2) + \text{In}(\boldsymbol{N}_3)$$

Thus, one is able to see how much of the total variation in the data is accounted for in each sub-matrix.

A description of s-CA as applied to a matrix $\boldsymbol{N}$, in the form of a contingency table, is presented in the succeeding text. Further details can be found in [5, 7, 10].

From the matrix $\boldsymbol{N}$ of non-negative numbers, the correspondence matrix, $\boldsymbol{P}$, is formed by dividing each element of $\boldsymbol{N}$ by its grand total. The elements of $\boldsymbol{P}$ can be thought of as the probability density of the cells of the matrix and the vectors of row and column sums of $\boldsymbol{P}$, denoted by **r** and **c**, as marginal densities. The elements of **r** and **c,** termed masses, are a measure of the relative importance of each row and column point. They are represented in diagonal matrices as $\boldsymbol{D_r}$ and $\boldsymbol{D_c}$ respectively. By dividing each element of a row (column) by its respective row (column) sum, we form a vector of relative frequencies that is called a row (column) profile. These profiles define the two clouds of points, one for rows and one for columns, in multi-dimensional weighted Euclidean space. The dimension weights for the row and column clouds are defined by the inverse of the elements of **c** $\left(\boldsymbol{D}_{\mathbf{c}}^{-1}\right)$ and **r** $\left(\boldsymbol{D}_{\mathbf{r}}^{-1}\right)$ respectively.

Under the assumption that the rows and columns of $\boldsymbol{P}$ are independent, the expected value of cell (i,j) of $\boldsymbol{P}$ is the product of the masses, $r_i c_j$. Calculating the difference between $p_{ij}$ and its expected value, $r_i c_j$, and then dividing by the square root of $r_i c_j$, serves to centre and normalise the correspondence matrix and results in a matrix of standardised residuals, which we shall call $\boldsymbol{S}$. The sum of squared elements of $\boldsymbol{S}$ is a measure of the total variation in the data and is termed total inertia.

It is at this stage that we 'interrupt' the CA process to implement the 'adjustment' needed for s-CA.

From the matrix, $\boldsymbol{S}$**,** of standardised residuals, select those rows and columns that make up the subset of variables/categories chosen to be included in further analysis. Let this matrix be $\boldsymbol{S}^*$. It is important to note that marginal densities, **r** and **c**, for the full matrix are retained for all future calculations [7].

The objective of CA and its variants, including s-CA, is to identify low-dimensional subspaces of the row and column clouds, which are closest to the points in terms of weighted sum of squared distances. This is achieved by performing an SVD on $\boldsymbol{S}^*$. In other words, $\boldsymbol{S}^* = \boldsymbol{U}\Delta\boldsymbol{V}^{\boldsymbol{T}}$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are the left and right singular vectors, respectively, and $\Delta$ is a diagonal matrix of singular values in decreasing order of magnitude. The principal axes of the row and column clouds are defined, respectively, by the $\boldsymbol{K}^*$ left and right singular vectors corresponding to the $\boldsymbol{K}^*$ largest singular values.

From the result of the SVD, we are able to define the principal coordinates of the points, that is, coordinates with respect to their principal axes. These row and column principal coordinates are calculated as $\boldsymbol{F} = \boldsymbol{D}_{\mathbf{r}}^{-1/2}\boldsymbol{U}\Delta$ and $\boldsymbol{G} = \boldsymbol{D}_{\mathbf{c}}^{-1/2}\boldsymbol{V}\Delta$, respectively. It is these coordinates that are used to produce the graphical displays of the points.

The amount of inertia explained by each principal axis is given by the square of the corresponding singular value.

### 2.2. *Chi-square test for independence*

Traditionally, the chi-square test for independence is used to test for significant associations between the rows and columns of a contingency table.

Each element of a random sample of size $n$ is classified according to two criteria. The first criteria, broken down into $r$ categories, and the second criteria, broken down into $c$ categories, represent the r rows and c columns of the table, respectively. The entry in the cell corresponding to row i and column j, $O_{ij}$, represents the number of elements that fall into that cell. The total number of observations in each row and column is found in the row and column sums R and C, respectively. The expected frequency, $E_{ij}$, in cell (i,j) equals $R_i C_j/n$. The chi-square test statistic is then calculated as

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Copyright © 2014 John Wiley & Sons, Ltd.

This commonly used statistic is a measure of how much the observed frequencies differ from what is expected.

Using this data set, the chi-square test was applied to contingency tables where the rows represented the asthma severity categories and the columns represented the categories of one of the environmental, genetic, behavioural and socio-economic variables. A significant result indicated that an association between asthma and the second variable did exist.

As an extension to the chi-square test, Cramer's V statistic was also calculated. This is defined as $V = \sqrt{\frac{\chi^2}{n(k-1)}}$, where $\chi^2$ and $n$ are as defined earlier and $k = \min(r, c)$. This gives a measure of the relative strength of association between two variables. A minimum threshold value of 0.1 suggests that there is a substantive relationship between the two variables.

## 3. The data

In 2004, the South Durban Health Study was commissioned by the eThekwini Municipality, South Africa, and undertaken by researchers at the University of KwaZulu-Natal. The objective of the study was to determine the effects of ambient pollution on childhood asthma, adjusting for covariates such as socio-economic status, allergy, genetic, environmental tobacco smoke exposure and other associated factors. Pollutants were monitored on a continuous basis throughout the study, while cross-sectional surveys and serial peak flow monitoring and symptoms logs provided health outcome information. Data for this study were gathered in the form of five surveys from children, their guardians and their families, at four primary schools in the south Durban area and three primary schools in the north Durban area. The data thus allow for comparisons to be made between data collected from the two areas [11].

At each school, in order to achieve a sample of persistent asthmatics with adequate power to determine association between asthma and the variables of interest, it was necessary for all students from Grades 3–6 to complete the 'screening' questionnaire, which contained questions regarding the child's respiratory health with specific reference to asthma and asthma symptoms. The study sample comprised all students from one randomly selected Grade 4 classroom and children with persistent asthma (on the basis of the screening questionnaire responses) from all Grades 3–6.

Trained interviewers from the research team administered the surveys. The 'caregiver', 'adult' and 'family' interviews were conducted with family members at home, while the 'child' interview was carried out at school.

From the 423 randomly selected and invited subjects in the study that formed the study sampling frame, 41 were excluded either because of the absence of asthma classification or inconsistencies across the five instruments for an individual participant. The final sample thus comprised 382 children.

Of the 382 subjects, 27 (7.1%) were classified as having moderate to severe asthma, 47 (12.3%) suffered from mild persistent asthma, 76 (19.9%) showed symptoms for mild intermittent (probable) asthma and the remaining 232 (60.7%) did not exhibit definite asthma symptoms. Classification into probable, mild persistent or moderate to severe asthma was based on the criteria provided by the US National Asthma Education Program [12].

Twenty one environmental, genetic, socio-economic and behavioural variables, broken down into 53 categories, and the four-tiered asthma severity variable were chosen from the different surveys to be used as variables in the analysis. The details of these, along with their frequencies, are presented in Table I.

As is the norm for survey-related data, there are many instances of non-response resulting in missing data. Of the 8404 possible data entries for the 22 variables, 445 (5.3%) are missing. There is a large non-response for 'income' where 19.4% of the respondents were reluctant to divulge their income. The missing items are confined to 166 (43.5%) of the 382 records, thus leaving a total of 216 complete records. Only four of the 22 variables are complete. Frequencies of non-response for each variable are included in Table I.

## 4. Application of the data set

### *4.1. Subset correspondence analysis*

The data, in the form of a contingency table, consist of four columns—representing the four asthma categories—and 71 rows—representing the categories of the 21 variables plus a separate missing category for each variable that suffered from non-response.

Copyright © 2014 John Wiley & Sons, Ltd.

**Table I.** Categories, code names and frequencies for all variables.

| Variables (survey*) | Categories (code names)—count ($N = 382$) | | | | | | | | Non-response—count (%) |
|---|---|---|---|---|---|---|---|---|---|
| Gender (All) | Male (male) | 163 | Female (fem) | 219 | | | | | 0 |
| Exercise (C) | <Twice weekly (E1) | 113 | 2–4 times/week (E2) | 135 | >4 times/week (E3) | 110 | | | E*—24 (6) |
| TV watching (C) | <1 h a day (T1) | 86 | 1–3 h/day (T2) | 193 | >3 h/day (T3) | 78 | | | T*—25 (7) |
| Smokers in the home (C) | Yes (SY) | 187 | No (SN) | 194 | | | | | Sm*—1 (<1) |
| Breakfast habits (C) | Daily (BD) | 236 | Not daily (BN) | 121 | | | | | B*—25 (7) |
| Pets at home (F/G) | Yes (PY) | 114 | No (PN) | 264 | | | | | P*—4 (1) |
| Food availability (F) | Enough food (FE) | 265 | Not enough (FN) | 85 | | | | | F*—32 (8) |
| Work and wear (F/G) | Yes (WWY) | 36 | No (WWN) | 332 | | | | | WW*—14 (4) |
| Smoke while pregnant (G) | Yes (SPY) | 35 | No (SPN) | 328 | | | | | SP*—19 (5) |
| Neonatal care (G) | Yes (NY) | 50 | No (NN) | 318 | | | | | N*—14 (4) |
| Birthweight (G) | Up to 2.5 kg (BW1) | 56 | >2.5 kg (BW2) | 280 | Do not know (BW?) | 42 | | | BW*—4 (1) |
| Fear in neighbourhood (C) | Yes (FrY) | 165 | No (FrN) | 192 | | | | | Fr*—25 (7) |
| Violence experienced (C) | Yes (VY) | 185 | No (VN) | 169 | | | | | V*—28 (7) |
| Attacked with weapons (C) | Yes (WY) | 160 | No (WN) | 194 | | | | | W*—28 (7) |
| Perceived weight (C) | Overweight (O) | 54 | Underweight (U) | 35 | Correct weight (C) | 267 | | | PW*—26 (7) |
| Smokers in cars (C) | Yes (SVY) | 94 | No (SVN) | 259 | | | | | SV*—29 (8) |
| Stove and fuel (F) | Paraffin stove (p) | 6 | Gas stove (g) | 3 | Electric stove (e) | 308 | No stove (n) | 27 | S*—38 (10) |
| Number of people in home (F) | 1–4 people (N1) | 124 | 5–7 people (N2) | 153 | >7 people (N3) | 70 | | | Np*—35 (9) |
| Age (C/G) | 8–9 years (A1) | 25 | 10 years (A2) | 196 | 11 years (A3) | 135 | 12+years (A4) | 26 | 0 |
| Income (F) | Up to R1000 (I1) | 79 | R1001–R4500 (I2) | 102 | R4501–R10000 (I3) | 88 | R100001+ (I4) | 39 | I*—74 (19) |
| Area (All) | south Durban (SD) | 197 | north Durban (ND) | 195 | | | | | 0 |
| Asthma severity (S) | Moderate/severe (ASMS) | 27 | Mild persistent (ASMP) | 47 | Mild intermittent (ASMI) | 76 | No asthma (ASN) | 232 | 0 |

*S, screening; G, caregiver; C, child; F, family.

Copyright © 2014 John Wiley & Sons, Ltd.

With the application of this data set, the objective was to identify relationships between the environmental, socio-economic, genetic and behavioural variables and to investigate possible relationships between these variables and asthma severity. CA was initially applied to the full data set. The total inertia amounted to 0.0207.

A number of the non-response categories contributed highly to the orientation of axis 2. This resulted in an elongation of the scale along this axis that, in turn, resulted in a clumping together of variables near the origin. This made it very difficult to distinguish between the points and interpret the maps, and masked more relevant relationships in the data. Furthermore, given the large number of variables in the data set, the inclusion of the non-response categories exacerbated the situation of an already crowded display. To address these phenomena, s-CA was applied to the subset of observed data, thus excluding the non-response categories from the analysis. The category BW?, a response option for respondents who did not know the birthweight, was also excluded as it was considered to play a similar role to BW* (non-response to birthweight question).

The total inertia accounted for by the subset of observed categories is 0.0162, which is 78.3% of the total inertia explained by the full data set.

*4.1.1. Interpretation of the principal axes.* The plots, in conjunction with the calculated contributions to inertia across the chosen dimensions (Table II), are used to identify and interpret the trends and relationships present in the data [5].

- For each principal axis, identify the largest values in the column headed CTR to interpret the dimensions. This enables us to assign 'meanings' to each axis. These values are scaled so that each column sums to 1000.
- For each point, examine the values in the COR columns across the dimensions to identify the axes that best represent the point. These values are a measure of how close a point lies to each of the axes and are independent of its mass or distance from the origin. High values of COR indicate that the axis contributes highly to the point's inertia; the angle the point makes with the axis will be small, and we can say that the point 'correlates' with the axis. Points with extremely high COR values are positioned nearly on the axis; this indicates that there is very little error in its location on the display.
- The values in the QLT column are calculated as the sum of the COR values across the dimensions. This is a measure of the quality of representation of the points in the subspace of chosen dimensionality. Values have been scaled so that, across all possible dimensions, QLT equals 1000.

We will interpret the first two axes that account for 88.92% of the total inertia. The total inertia is an indication of the accuracy of the display. Thus, in this example, we have 11.08% error in the display. Equivalently, the two-dimensional figure accounts for 88.92% of the variability in the data, which leaves 11.08% unaccounted for.

Axis 1—the variables that make the most contribution to the orientation of this axis are A1 (age 8—9 years) and NY (having received some form of special neonatal care). Both are physiological variables, and they have been separated out from the other variables and are situated on the negative side of the axis. Other variables that have contributed to this axis and are associated with the aforementioned variables are WWY (exposure to secondary smoke and chemicals), male, N1 (up to four people in the home), I3 (income of R4501–R10000), T1 (<1 h TV a day), SD (from south Durban), p (those who use a paraffin stove) and BW1 (<2.5 kg at birth). Opposing these, on the positive side, are PY (having had a pet), ND (from north Durban) and female. Moderate to severe (ASMS) and mild persistent (ASMP) asthma are associated with the groupings on the negative side and 'no asthma' (ASN) with the group on the positive side.

Many variables have not played a major role in the orientation of the axis but are correlated with it, as evidenced by the large COR values. In particular, the smoke exposure variables, both in the home (SY) and in vehicles (SVY), are highly correlated with this axis and are situated on the negative side indicating an association with the more severe levels of asthma.

It is evident that subjects are separated on this axis on the basis of both physiographic factors and smoke exposure. These are the biggest contrasts in the data and account for 66.52% of the total inertia.

Axis 2—the orientation of this axis is defined mainly by the variables I1 (income of <R1000), male and female, T1 ( less than 1 h TV a day) and WY (being attacked with weapons). There is a separation on this axis of those subjects who are from the lowest income group (I1), are male, experience fear in the neighbourhood (FrY), have been attacked with weapons (WY) and watch TV for less than an hour

Copyright © 2014 John Wiley & Sons, Ltd.

**Table II.** Decomposition of inertia for the first three principal axes.

| Name | Mass | QLT | INR | $k=1$ | COR | CTR | $k=2$ | COR | CTR |
|---|---|---|---|---|---|---|---|---|---|
| A1 | 3 | 944 | 3 | −717 | 943 | 149 | −28 | 1 | 1 |
| A2 | 24 | 719 | 0 | 34 | 712 | 3 | 3 | 7 | 0 |
| A3 | 17 | 750 | 0 | 73 | 725 | 8 | −13 | 25 | 1 |
| A4 | 3 | 154 | 0 | 57 | 59 | 1 | 72 | 95 | 5 |
| male | 20 | 881 | 2 | −149 | 455 | 42 | 144 | 426 | 116 |
| fem | 27 | 881 | 1 | 111 | 455 | 31 | −107 | 426 | 86 |
| BW1 | 7 | 997 | 1 | −220 | 673 | 31 | −153 | 324 | 45 |
| BW2 | 35 | 907 | 0 | 37 | 350 | 4 | 46 | 557 | 21 |
| NY | 6 | 974 | 0 | −475 | 974 | 131 | −10 | 0 | 0 |
| NN | 40 | 958 | 68 | 65 | 882 | 16 | 19 | 76 | 4 |
| FrY | 21 | 974 | 18 | 68 | 395 | 9 | 83 | 579 | 39 |
| FrN | 24 | 944 | 19 | −46 | 370 | 5 | −58 | 574 | 22 |
| SPY | 4 | 257 | 3 | 13 | 28 | 0 | −37 | 229 | 2 |
| SPN | 41 | 986 | 18 | −7 | 152 | 0 | 16 | 834 | 3 |
| SY | 23 | 998 | 0 | −51 | 862 | 6 | −20 | 136 | 3 |
| SN | 24 | 991 | 32 | 47 | 816 | 5 | 22 | 175 | 3 |
| SVY | 12 | 985 | 8 | −76 | 975 | 6 | −8 | 10 | 0 |
| SVN | 32 | 989 | 4 | 30 | 870 | 3 | 11 | 119 | 1 |
| E1 | 14 | 737 | 17 | −2 | 2 | 0 | −29 | 735 | 3 |
| E2 | 17 | 824 | 5 | 60 | 666 | 6 | −29 | 158 | 4 |
| E3 | 14 | 934 | 14 | −15 | 41 | 0 | 71 | 893 | 19 |
| T1 | 11 | 751 | 18 | −187 | 445 | 35 | 155 | 306 | 71 |
| T2 | 24 | 634 | 11 | 45 | 326 | 5 | −44 | 308 | 13 |
| T3 | 10 | 999 | 92 | 147 | 985 | 19 | −18 | 14 | 1 |
| N1 | 15 | 986 | 16 | −169 | 890 | 41 | −56 | 96 | 13 |
| N2 | 19 | 940 | 2 | 61 | 634 | 7 | 42 | 306 | 9 |
| N3 | 9 | 820 | 14 | 161 | 779 | 21 | −37 | 41 | 3 |
| I1 | 10 | 722 | 3 | 38 | 21 | 1 | 221 | 701 | 132 |
| I2 | 13 | 951 | 13 | 34 | 118 | 1 | −91 | 833 | 29 |
| I3 | 11 | 836 | 1 | −189 | 612 | 36 | −114 | 224 | 40 |
| I4 | 5 | 970 | 134 | 102 | 970 | 5 | 2 | 0 | 0 |
| FN | 33 | 763 | 29 | 59 | 646 | 11 | −25 | 117 | 6 |
| FE | 11 | 989 | 9 | −32 | 150 | 1 | 75 | 839 | 17 |
| O | 7 | 914 | 22 | 169 | 832 | 18 | −53 | 82 | 5 |
| C | 4 | 814 | 88 | −187 | 363 | 14 | 208 | 451 | 52 |
| U | 33 | 260 | 32 | −3 | 9 | 0 | −13 | 251 | 2 |
| WWY | 4 | 928 | 0 | −406 | 911 | 69 | −55 | 17 | 4 |
| WWN | 41 | 944 | 42 | 35 | 663 | 5 | 23 | 281 | 6 |
| PY | 14 | 816 | 0 | 200 | 739 | 53 | −65 | 77 | 16 |
| PN | 33 | 887 | 26 | −80 | 702 | 20 | 41 | 185 | 15 |
| SD | 23 | 932 | 26 | −126 | 845 | 34 | 40 | 87 | 10 |
| ND | 24 | 932 | 16 | 120 | 845 | 33 | −39 | 87 | 10 |
| BN | 29 | 952 | 1 | −19 | 377 | 1 | −23 | 575 | 4 |
| BD | 15 | 766 | 2 | 71 | 468 | 7 | 56 | 298 | 13 |
| VY | 23 | 994 | 28 | 111 | 912 | 26 | 33 | 82 | 7 |
| VN | 21 | 999 | 17 | −95 | 945 | 18 | −23 | 54 | 3 |
| WY | 20 | 995 | 22 | 99 | 492 | 18 | 100 | 503 | 55 |
| WN | 24 | 985 | 6 | −50 | 358 | 6 | −66 | 627 | 29 |
| p | 1 | 983 | 15 | −689 | 671 | 33 | −470 | 312 | 45 |
| g | 0 | 999 | 96 | 453 | 707 | 7 | 291 | 292 | 9 |
| e | 38 | 174 | 1 | 14 | 173 | 1 | 1 | 1 | 0 |
| n | 3 | 788 | 2 | −55 | 451 | 1 | −47 | 337 | 2 |
| ASMS | 71 | 951 | 85 | −312 | 909 | 638 | 67 | 42 | 88 |
| ASMP | 123 | 777 | 148 | −131 | 563 | 195 | −80 | 214 | 220 |
| ASMI | 199 | 887 | 197 | 45 | 142 | 38 | 103 | 745 | 585 |
| ASN | 607 | 865 | 570 | 48 | 676 | 130 | −25 | 189 | 108 |

Mass (Mass) and inertia (INR) of each variable; the quality (QLT) of the variable's representation in the subspace of the first two axes; coordinates (k = ...); contributions of axes to the inertia of the variables (COR); and contributions of variables to the inertia of the axes (CTR)*.

* For details of the formulae for calculations, see [5], p. 91.

Copyright © 2014 John Wiley & Sons, Ltd.

a day (T1) from those subjects who are female, have not been attacked with weapons (WN), are from the R1001–R4500 income group (I2) and do not experience fear in the neighbourhood (FrN). The mild intermittent asthma variable (ASMI) correlates with the former grouping. Axis 2 can be thought of as distinguishing between subjects on the basis of their socio-economic status(SES) and accounts for 22.4% of the total inertia.

*4.1.2. Graphical displays.* In the graphical display, those variables that are not well represented in the subspace are situated near the origin and do not add to the interpretation of the display. By examining the angles that the points make with each other and with the principal axes, we can identify and interpret trends and relationships present in the data.

In the plane of the first and second axes (Figure 1), which accounts for 88.9% of the variation in the data, the physiological/smoke exposure axis is plotted against the socio-economic axis. Variables indicative of low socio-economic status are situated above the horizontal axis and the higher socio-economic variables below. In the same way, the vertical axis separates the smoke exposure variables as well as those representing low birthweight (BW1), having had neonatal care (NY), male and low age (A1) from their 'opposites'. The asthma variables are well represented in this subspace. The more severe asthma variables (ASMS and ASMP) are split from the other categories (ASMI and ASN) by the vertical axis indicating an association of worse asthma with those variables situated to the left of the axis. Mild intermittent asthma (ASMI) is removed from the other three asthma variables and tends in the direction of lower socio-economic status. Further distinctions between the levels of asthma severity are evidenced by their locations—each in a different quadrant. The strongest associations with moderate to severe asthma (ASMS) were shown by men, having had neonatal care (NY), smoke exposure in vehicles (SVY), 8- to 9-year olds (A1) and coming from south Durban (SD); mild persistent asthma was associated most with a birthweight of less than 2.5 kg (BW1), using a paraffin stove (p) or not having a stove (n), smoke exposure in the home (SY), exposure to secondary smoke and chemicals (WWY), living in a home with up to four people (N1) and a monthly income of R4501–R10000 (I3); and associations with mild intermittent asthma were shown by the lowest income group (I1), a birthweight of more than 2.5 kg (BW2), being attacked by weapons (WY), experiencing fear in the neighbourhood (FrY) and doing exercise more than four times a week (E3).

An interesting result is the distinction between the different forms of smoke exposure and their associations with asthma severity. A close association is evident between smoke exposure in the home (SY) and mild persistent asthma (ASMP). Smoke exposure in a vehicle (SVY) shows a stronger association with moderate to severe asthma (ASMS) than with mild persistent asthma (ASMP), as indicated by the angles that the point vectors make with the asthma variables. Exposure to severe levels of air pollution, as experienced in the south Durban region (SD), shows a strong association with moderate to severe

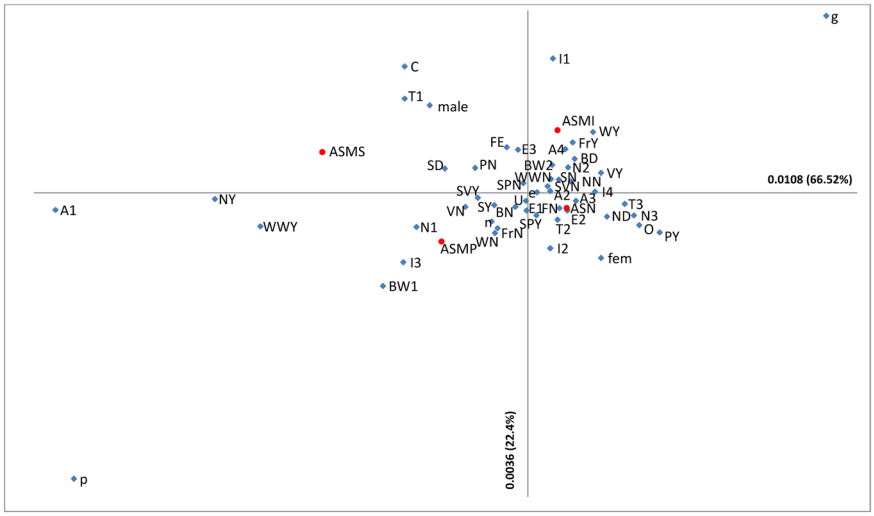



**Figure 1.** Subset correspondence analysis map of contingency table with the row points represented by ♦ and the column points by • projected onto the plane of the first and second principal axes. Values on the axes indicate principal inertias and their respective percentages of total inertia.

Copyright © 2014 John Wiley & Sons, Ltd.

**Table III.** Results of Pearson's chi-square and Cramer's V tests for the 10 variables that exhibit the strongest relationship with asthma severity.

| Variable (categories) | Chi-square *p*-value | Cramer's V |
|---|---|---|
| Gender (male/female) | 0.003 | 0.190 |
| Neonatal care (NY/NN)) | 0.005 | 0.186 |
| Pets (PY/PN) | 0.036 | 0.150 |
| Work and wear (WWY/WWN) | 0.070 | 0.138 |
| Area (SD/ND) | 0.077 | 0.134 |
| Age (A1/A2/A3/A4) | 0.087 | 0.120 |
| TV (T1/T2/T3) | 0.137 | 0.117 |
| Birthweight (BW1/BW2) | 0.182 | 0.120 |
| Income (I1/I2/I3/I4) | 0.216 | 0.114 |
| Weapons (WY/WN) | 0.217 | 0.112 |

asthma (ASMS). Smoking while pregnant (SPY) is not well represented in this subspace and is therefore not included in this discussion.

Another interesting phenomenon is the positioning of the stove variables paraffin (p) and gas (g) at opposite corners of the display. The association of gas stove (g) with mild intermittent asthma (ASMS) contrasts that of paraffin stove (p) with mild persistent asthma (ASMP).

With regard to the number of people in the home and its association with asthma severity, results show that N1 (1–4 people) tends in the direction of mild persistent asthma (ASMP), N2 (5–7 people) tends towards mild intermittent asthma (ASMI) and N3 (8+ people) tends towards no asthma (ASN). Thus, the fewer people there are in the home, the higher the level of asthma severity.

It can be seen that the inertia associated with this subspace amounts to 0.0144 (0.0108 + 0.0036) in total. This relatively low value indicates that there is not a lot of variability in the data and explains the bunching up of the variables in the display [10].

### *4.2. Chi-square analysis*

As a comparative method of association analysis of contingency tables, Pearson's chi-square test was applied to individual cross-tabulations of asthma severity with each of the generic, socio-economic, behavioural and environmental variables. By examining the contributions of individual cells to the chi-square value, we were able to identify specific relationships between the two variables in the table. In addition, Cramer's V statistic gave us an indication of the relative strength of the associations found.

Results from Pearson's chi-square (Table III) showed that there was agreement, at the 5% level of significance, that gender, neonatal care and ever having pets are significantly related to asthma severity. Specifically, significantly more than expected of the subjects who were male or who had had specialist care at birth and significantly fewer than expected of those who ever had pets suffered from moderate to severe asthma. Relaxing the level of significance to 10%, associations were found to exist between asthma severity and age, area and exposure to secondary chemicals and dust. More specifically, more than expected of the youngest age group as well as those who were exposed to secondary chemicals and smoke suffered from moderate to severe or mild persistent asthma, while more than expected of those from south Durban had moderate to severe asthma. Cramer's V statistic (Table III) indicates that the three strongest associations are exhibited between asthma severity and gender, neonatal care and pets, respectively. While the values of this statistic signify only a low association for each of the variables shown, they are large enough to suggest that a relationship between asthma severity and each of these variables does exist.

## 5. Discussion

In our application of s-CA to a data set with a substantial amount of missing data, we were able to show that the use of this technique provides a meaningful approach to exploring the relationships between categorical variables that suffer from missingness. This approach provides several advantages when compared with other methods of addressing such shortcomings of data sets. The advantages are that the method is not constrained by either model assumptions or distributional requirements, it can be applied

Copyright © 2014 John Wiley & Sons, Ltd.

irrespective of the missingness mechanism present, it is computationally simple and it is able to handle large numbers of categorical variables. All the standard analyses were performed using SPSS (version 17), and a macro program was written to perform the s-CA.

Applying CA to the full data set resulted in an elongation of the scale on axis 2, which exacerbated an already crowded display, thus making it difficult to identify points and interpret relationships between them. In addition, it is the relationships between the measured variables and level of asthma severity that are of interest in this study. Because of the useful property of s-CA, whereby the full data matrix can be partitioned into smaller mutually exclusive sub-matrices, with the respective decomposition of the total inertia, CA was applied to the sub-matrix of observed variable categories only, which allowed for a clearer display of the points and enabled the exploration of the relationships between the relevant variables.

The application of this novel explorative statistical technique has enabled us to examine a large number of environmental, behavioural, genetic and socio-economic variables to uncover relationships between these variables and, at the same time, retain all records. Furthermore, associations between these variables and asthma have been found that generally confirm established theories regarding factors that exacerbate asthma. We have further been able to distinguish between different levels of asthma severity and the factors that are associated with them.

There is agreement that asthma is associated with the following: younger children [13]; a birthweight of less than 2.5 kg and having had neonatal care [14]; exposure to low concentrations of compounds and pollutants as a result of living in the same house with someone who works in a chemical/dust environment and wears their work clothes ate home [15, 16]; male children [17, 18]; and smoke exposure both in vehicles [19], in the home [20, 21] and in the form of air pollution [22, 23]. These variables are shown to be associated with the higher levels of asthma severity in this application.

Other studies that have lead to results that confirm documented theories for factors that influence asthma severity include the following: that the risk from exposure to smoke in a car smoke exceeds the risk from smoke in the home [24]; that there is an association between asthma and indicators of low SES, viz. experiencing fear in the neighbourhood [25], neighbourhood stressors in the form of the use of weapons [26, 27] and low income homes [28, 29]; and that asthma occurrence is inversely related to the size of the family [30].

Relative weights and inter-point distances are retained from the analysis of the full data set and are not recalculated for the analysis of the subset. This allows for the decomposition of the inertia into parts representing mutually exclusive and exhaustive subsets. CA of the full data set resulted in a total inertia of 0.0207. This is a measure of the dispersion of the points in the full m-dimensional space. The analysis of the subset of observed categories yielded a total inertia of 0.0162, and total inertia from the analysis of the non-response categories is 0.0045. Because the two subsets are mutually exclusive and exhaustive, the sums of their total inertias equal the total inertia of the whole data set. Furthermore, the observed categories account for nearly four times as much of the inertia ($0.0162/0.0207 = 78.3\%$) in the data as is attributed to the non-response categories ($0.0045/0.0207 = 21.7\%$). While we have been able to identify many interesting relationships in the data, we can see from the correspondence map that the dispersion of the points is not extensive. This is borne out by the value of the total inertia (a relatively low 0.0162), which is a measure of how much the measured profiles are spread around the origin.

While it is important to note that, with s-CA, relationships found to exist between variables/categories cannot be assumed to be statistically significant, comparative tests of association were carried out on cross-tabulations of asthma severity with the other variables. Relationships between asthma severity and a number of the variables included in the study were identified. Despite the fact that the associations were not necessarily strong, they do corroborate the associations found with s-CA. The fact that only a few variables were found to be significantly associated with asthma severity is consistent with our finding in s-CA that the dispersion of points was not large, as seen both in the graphical display and in the low inertia value.

We thus have shown that s-CA, as presented here, has a two-fold purpose: firstly, as an exploratory tool to seek interrelationships between variable categories and to identify those variable categories that are associated with different levels of childhood asthma so that they can be taken further and used in more rigid analysis, and secondly, to manage the missing data and the problem of crowding created by it. Furthermore, where large numbers of variables/categories are involved, relationships between variables/categories are not generally easy to summarise. So, we could take this a step further and suggest subsequent division of the data into numerous smaller, sensibly selected, mutually exclusive and exhaustive subsets. In these situations, we thus propose that s-CA is an ideal choice of method and

Copyright © 2014 John Wiley & Sons, Ltd.

produces easily interpreted graphical output to provide a general view of the associations between the many variables.

In conclusion, despite the presence of missing data, s-CA is able to explore the data as a whole and represent the variables graphically, thus implying relationships between variables. By identifying those variables important to the determination of the principal axes, the identification of a selection of the variables to take forward for further analysis is possible. We believe that our exploratory method is easier to apply than the existing multiple imputation methods in which many complexities need to be considered. While multiple imputation allows one to carry out statistical analysis on data that encounters missingness, the sophistications in the assumptions about the model, the missingness mechanisms and the computational algorithms are restrictive and make it more difficult to use. We hope that the s-CA approach will offer an alternative paradigm to dealing with the analysis of categorical data that suffer from missingness.

## Acknowledgements

This work was supported by eThekwini Metropolitan Municipality (local government) contract no 1A-103, Medical Research Council of South Africa and University of KwaZulu-Natal Research Funds. We are grateful to the editor, Dr Louise Ryan, Dr Michael Greenacre and other referees for their detailed comments and suggestions that greatly helped to improve the paper. Thanks also go to Dr Graciella Mentz for her part in the earlier stages of the project with the data collection and cleaning.

## References

1. Eekhout I, de Boer RM, Twisk JW, de Vet HC, Heymans MW. Missing data: a systematic review of how they are reported and handled. *Epidemiology* 2012; **23**:729–732. DOI: 10.1097/EDE.0b013e3182576cdb.
2. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 1987.
3. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology* 1995; **142**:1255–1264.
4. Greenacre MJ, Blasius J. Correspondence analysis and related methods in practice. In *Correspondence Analysis and Related Methods*, Greenacre M, Blasius J (eds). Chapman & Hall/CRC: Boca Raton, 2006; 3–40.
5. Greenacre MJ. *Theory and Applications of Correspondence Analysis*. Academic Press: London, 1984.
6. Greenacre M, Pardo R. Subset correspondence analysis visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods & Research* 2006; **35**:193–218.
7. Greenacre MJ, Pardo R. Multiple correspondence analysis of subsets of response categories. In *Correspondence Analysis and Related Methods*, Greenacre M, Blasius J (eds). Chapman & Hall/CRC: Boca Raton, 2006; 197–217.
8. Greenacre MJ. Some objective methods of graphical display of a data matrix. *Special Report*, Department of Statistics and Operations Research, University of South Africa, 1978.
9. Greenacre MJ, Pardo Avellaneda R. Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey, 2005.
10. Greenacre M. Correspondence analysis in medical research. *Statistical Methods in Medical Research* 1992; **1**:97–117.
11. Naidoo RN, Robins T, Batterman S, Mentz G, Jack C. Ambient pollution and respiratory outcomes among schoolchildren in Durban, South Africa. *South African Journal of Child Health* 2013; **7**:127–134.
12. Program NAEP. Expert panel report: guidelines for the diagnosis and management of asthma. *In Expert Panel Report: Guidelines for the Diagnosis and Management of Asthma*, US Department of Health and Human Services: City, 1991.
13. Asher MI, Montefort S, Björkstén B, Lai C, Strachan DP, Weiland SK, Williams H. ISAAC Phase Three Study Group: worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys. *Lancet* 2006; **368**:733–743.
14. Mai XM, Gäddlin PO, Nilsson L, Finnström O, Björkstén B, Jenmalm MC, Leijon I. Asthma, lung function and allergy in 12-year-old children with very low birth weight: a prospective study. *Pediatric Allergy and Immunology* 2003; **14**:184–192. DOI: 10.1034/j.1399-3038.2003.00045.x.
15. Becher R, Hongslo JK, Jantunen MJ, Dybing E. Environmental chemicals relevant for respiratory hypersensitivity: the indoor environment. *Toxicology Letters* 1996; **86**:155–162. DOI: 10.1016/0378-4274(96)03685-5.
16. Venables KM, Chan-Yeung M. Occupational asthma. *Lancet* 1997; **349**:1465–1469. DOI: 10.1016/S0140-6736(96)07219-4.
17. Bonner J. The epidemiology and natural history of asthma. *Clinics in Chest Medicine* 1984; **5**:557–565.
18. Almqvist C, Worm M, Leynaert B. Impact of gender on asthma in childhood and adolescence: a GA2LEN review. *Allergy* 2007; **63**:47–57. DOI: 10.1111/j.1398-9995.2007.01524.x.
19. Sendzik T, Fong GT, Travers MJ, Hyland A. An experimental investigation of tobacco smoke pollution in cars. *Nicotine & Tobacco Research* 2009; **11**:627–634. DOI: 10.1093/ntr/ntp019.
20. Charoenca N, Kungskulniti N, Tipayamongkholgul M, Sujirarat D, Lohchindarat S, Mock J, Hamann SL. Determining the burden of secondhand smoke exposure on the respiratory health of Thai children. *Tobacco Induced Diseases* 2013; **11**:7.
21. Ehrlich R, Kattan M, Godbold J, Saltzberg DS, Grimm KT, Landrigan P, Lilienfeld D. Childhood asthma and passive smoking. *American Review of Respiratory Disease* 1992; **145**:594–599.

Copyright © 2014 John Wiley & Sons, Ltd.

22. Neidell MJ. Air pollution, health, and socio-economic status: the effect of outdoor air quality on childhood asthma. *Journal of Health Economics* 2004; **23**:1209–1236. DOI: 10.1016/j.jhealeco.2004.05.002.
23. Peden DB. The epidemiology and genetics of asthma risk associated with air pollution. *Journal of Allergy and Clinical Immunology* 2005; **115**:213–219. DOI: 10.1016/j.jaci.2004.12.003.
24. Sly PD, Deverell M, Kusel MM, Holt PG. Exposure to environmental tobacco smoke in cars increases the risk of persistent wheeze in adolescents. *Medical Journal of Australia* 2007; **186**(6):322.
25. Subramanian S, Kennedy MH. Perception of neighborhood safety and reported childhood lifetime asthma in the United States (US): a study based on a national survey. *PloS One* 2009; **4**:e6091. DOI: 10.1371/journal.pone.0006091.
26. Wright RJ, Mitchell H, Visness CM, Cohen S, Stout J, Evans R, Gold DR. Community violence and asthma morbidity: the Inner-City Asthma Study. *American Journal of Public Health* 2004; **94**:625–632. DOI: 10.2105/AJPH.94.4.625.
27. Jeffrey J, Sternfeld I, Tager I. The association between childhood asthma and community violence, Los Angeles County, 2000. *Public Health Reports* 2006; **121**:720–728.
28. Cesaroni G, Farchi S, Davoli M, Forastiere F, Perucci C. Individual and area-based indicators of socioeconomic status and childhood asthma. *European Respiratory Journal* 2003; **22**:619–624. DOI: 10.1183/09031936.03.00091202.
29. Poyser M, Nelson H, Ehrlich R, Bateman E, Parnell S, Puterman A, Weinberg E. Socioeconomic deprivation and asthma prevalence and severity in young adolescents. *European Respiratory Journal* 2002; **19**:892–898. DOI: 10.1183/09031936.02.00238402.
30. Matricardi PM, Franzinelli F, Franco A, Caprio G, Murru F, Cioffi D, Ferrignoc L, Palermoa A, Ciccarelli N, Rosmini F. Sibship size, birth order, and atopy in 11,371 Italian young men. *Journal of Allergy and Clinical Immunology* 1998; **101**:439–444.

Copyright © 2014 John Wiley & Sons, Ltd.

Hendry et al. BMC Medical Research Methodology 2014, 14:136
http://www.biomedcentral.com/1471-2288/14/136

BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

# Model development including interactions with multiple imputed data

Gillian M Hendry[1*], Rajen N Naidoo[2], Temesgen Zewotir[1], Delia North[1] and Graciela Mentz[3]

## Abstract

**Background:** Multiple imputation is a reliable tool to deal with missing data and is becoming increasingly popular in biostatistics. However, building a model with interactions that are not specified *a priori*, in the presence of missing data, presents a challenge. On the one hand, the interactions are needed to impute the data, while on the other hand, the data is needed to identify the interactions. The objective of this study was to present a way in which this challenge can be addressed.

**Methods:** This paper investigates two strategies in which model development with interactions is achieved using a single data set generated from the Expectation Maximization (EM) algorithm. Imputation using both the fully conditional specification approach and the multivariate normal approach is carried out and results are compared. The strategies are illustrated with data from a study of ambient pollution and childhood asthma in Durban, South Africa.

**Results:** The different approaches to model building and imputation yielded similar results despite the data being mainly categorical. Both strategies investigated for building the model using the multivariate normal imputed data resulted in the identical set of variables and interactions being identified; while models built using data imputed by fully conditional specification were marginally different for the two strategies. It was found that, for both imputation approaches, model building with backward elimination applied to the initial EM data set was easier to implement, and produced good results, compared to those from a complete case analysis.

**Conclusions:** Developing a predictive model including interactions with data that suffers from missingness is easily done by identifying significant interactions and then applying backward elimination to a single data set imputed from the EM algorithm. It is hoped that this idea can be further developed and, by addressing this practical dilemma, there will be increased adoption of multiple imputation in medical research when data suffers from missingness.

**Keywords:** Interactions, Missing data, Model development, Multiple imputation, Ordinal regression

## Background

It is not unusual to encounter missing data in epidemiological studies [1,2]. Its presence affects the analysis of the data, and the methods employed in handling missing data can affect the results of the analysis. This could compromise conclusions drawn from the results. Types of missingness have been well documented [3]. Popular classifications are "missing completely at random" (MCAR – the missing values are independent of both observed and unobserved data); "missing at random" (MAR – the missing values are independent of unobserved data but may depend on observed data) and "not missing at random" (MNAR – the missing data depends on both observed and unobserved data).

Commonly, missing data is managed by simply dropping all cases that are not fully measured. However, such a complete case analysis can introduce bias into the results and, in some cases, wrong conclusions can be drawn [4]. While this approach is acceptable when the incomplete cases do not exceed 5% [5] and for which the missingness can be classified as MCAR, when these conditions are not met, alternative means of dealing with

* Correspondence: hendryfam@telkomsa.net
[1]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville Campus, University Road, Westville, Durban, South Africa
Full list of author information is available at the end of the article

BioMed Central

© 2014 Hendry et al.; licensee BioMed Central. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

the missing data need to be considered. One such method that is increasingly being used is multiple imputation (MI) [6].

Imputation of missing data on a variable involves replacing the missing value by a value drawn from an estimate of the distribution of the variable [7]. Multiple imputation does not replace the missing item with a single predicted value, but rather imputes multiple values for each missing data item. These multiple imputations and the addition of random error to each imputed item ensures that the variation in the imputed values follows closer the true distribution of the original measure. Multiple imputation is successfully applied to data that is MAR and yields unbiased results with accurate estimates for the standard errors [7]. Unfortunately, the missingness mechanism is not usually fully known and is often a combination of more than one mechanism. However, by ensuring that the imputation model is more general than the analysis model, multiple imputation will usually produce sound results [8-11]. This is achieved by including, in the imputation model, variables that are related to the incomplete variables as well as those related to their missingness; the outcome variable; and all interactions that will be examined in the analysis.

Rubin [12] suggests that the need to include all possibly relevant predictors in the imputation model is demanding in practice. If interactions are selected *a priori*, it is a straightforward exercise to include them in the imputation model [9]. If, on the other hand, the relevant interactions have not been identified beforehand, then ideally all possible interactions should be included in the imputation model. This is neither practical nor, in some cases, possible [13,14], particularly when the number of variables is large. While model development with multiple imputation has been documented [13,15-17], none of these studies addresses the issue of how to include, in the imputation model, interactions that are not known *a priori*. Developing a model with many variables, in the presence of missing data, when predictor variables include not only main effects but also interactions that are not pre-selected, presents a challenge, and not extensively reported in the literature. On the one hand, the data is needed to identify relevant interactions; on the other hand, the interactions are needed to impute the data. This paper addresses this dilemma and suggests a method in which model development, including interactions, and analysis can be carried out when missing data is imputed using multiple imputation.

We propose to identify the relevant interactions using a single complete set of data generated using the expectation-maximization (EM) algorithm for covariance matrices and then include these interactions in the imputation model.

## Methods

### The data

The relationship between environmental, socio-economic and genetic factors and the respiratory health of children in the Durban South region of KwaZulu-Natal, South Africa using cross-sectional data was investigated. The data comes from research commissioned by the eThekwini Municipality, Durban, South Africa in 2004 to investigate possible causal effects of environmental and lifestyle factors on respiratory health in children [18]. Ethical approval was obtained from the Ethics Committee of the University of KwaZulu-Natal (Ref No.: E117/03). All the legal guardians of the child participants in this study gave written informed consent, participated voluntarily, and had the right to withdraw at any stage.

After an asthma symptoms screening survey, a sample of 423 primary school children were invited to participate in the study and from each participant multiple questionnaires were required to be completed. Of the 423 children included in the study, 382 that were deemed to have reliable data as well as complete data on the outcome variable, asthma severity, were used for this analysis. The removal of these children did not result in any selection bias.

Most of the predictor variables suffered from missing data. A study on the missingness mechanism was made prior to imputing the missing values. For each incomplete variable, an indicator variable was created and Chi-square analyses were performed to test whether either the incomplete variable or its missingness was related to observed values of other variables.

### Selection of interactions for the imputation model

In order to ensure that the imputation model is at least as complex as the analysis model, and that the assumption of MAR is plausible, it is necessary to include the outcome variable and all possible likely predictors for the analysis model, in the imputation model. The selection of the interaction terms presents difficulties [16,17]. Comparable to the suggestion made by White et al [16], we have generated a single complete set of data using the EM algorithm for covariance matrices. The EM algorithm is an iterative procedure that can be used to create a complete data set in which all missing values are replaced by maximum likelihood (ML) values that are asymptotically unbiased. The process starts by replacing each missing value with an estimate calculated from a regression equation in which all the other variables are predictors. Once all the missing values have been replaced, a variance covariance matrix and a vector of means from the completed data are calculated. New regression equations are then formed to predict a new set of estimates for the missing values. This process is repeated until the variances, covariances and means converge, thus producing ML estimates of the parameters.

The complete data set generated from this process is then used for model development and the identification of interactions. In our application, convergence was achieved in 36 iterations.

### Multiple imputation

The imputation of multiple data sets was carried out using two different algorithms – multivariate normal imputation (MVNI) and fully conditional specification (FCS).

MVNI – This imputation algorithm, adopted by the NORM software [19], assumes the complete data (observed and missing values) follows a multivariate normal distribution. NORM uses a data augmentation (DA) procedure to impute multiple sets of data.

This two-step process makes use of the ML estimates from EM as parameter starting values. In the first step, DA randomly imputes the missing data using the assumed values of the parameters. In the second step, new parameter estimates are drawn from a Bayesian posterior distribution based on the observed and imputed data. The repetition of these two steps results in a Markov chain. DA converges when the distribution of parameter estimates stabilizes. Research has shown that DA nearly always converges in fewer cycles than does EM [8]. This enables one to estimate the cycle length, k, of DA as being any number at least as large as the number of iterations needed for EM to converge.

In order to impute m sets of data, DA is run for N = mk iterations and the data set at the end of every $k^{th}$ cycle is saved.

Because the data contained categorical variables, some adjustments were necessary both before and after imputation. Before imputation, dummy coding was applied to all the categorical variables and interaction product terms with more than two categories. After imputation, sensible rounding [20] was used on these variables to prepare the data for analysis.

FCS – FCS, also termed "chained equations", is the multiple imputation algorithm adopted by SPSS [21]. This is a more flexible approach to imputation in that it is designed to handle different types of variables (continuous, binary, categorical, ordinal) and does not assume multivariate normality of the data [6].

In practice, FCS involves running a series of regression models such that each variable with missing data is regressed on the other variables in the data set according to its distribution. So, for example, categorical variables will be modelled using logistic regression and continuous variables will be modelled using linear regression.

Imputation by FCS, as applied in SPSS, is also an iterative process that starts by imputing every missing value with random draws from the distribution of the non-missing values. Continuous variables are replaced with draws from a normal distribution and categorical variables are replaced with draws from a multinomial distribution. Azur et al [22] refer to these replacements as "place holders".

Each iteration involves the following steps:

- Set the "place holders" of one variable that suffers from missing values back to missing
- Set up a regression equation, according to the distribution of the variable, with the observed values as the dependent variable and the other variables as independent variables
- Replace the missing values from this variable with predictions from the regression equation
- Repeat these steps for each variable that has missing values.

This forms one iteration of the process. At each iteration the imputed values are updated. This process is repeated for a specified number of iterations, n, after which the data set is retained as one complete imputed data set. The number of iterations, n, chosen so that the parameters from the regression models have stabilized, is generally about ten [23]. This entire process is repeated until the required number, m, of imputed data sets is generated.

Each of the m data sets were analysed with ordinal regression – the chosen method of analysis – and the results were combined using Rubin's rules [4]. Although, in the past, it was widely thought that as few as 3 imputed data sets are needed to obtain good results and inferences, new studies have shown that this may, in fact, not be enough [24]. Studies have shown that there could be an important reduction in statistical power if m is small [9]. Graham et al [24] completed a simulation study on the number of imputations needed to attain maximum power. Their recommendations for the number of imputations, m, as a function of the fraction of missing information are summarized in Table 1. On the basis of the percentage of data missing in this study (5.3%), 20 sets of data were imputed.

### Model development

In order to develop the best model given the large number of variables available, the following three-stage process was followed. Firstly, all variables were purposefully selected as main effects. Secondly, in developing the full model, interactions were chosen one at a time in a stepwise manner such that the interaction that made

**Table 1 Recommended number of imputations needed for varying fractions of missing data (Graham [9])**

| Fraction of missing data | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|---|
| Number of imputations | 20 | 20 | 40 | 100 | >100 |

the biggest significant improvement to the fit was added to the model. For this process a cut-off $p$-value of 0.05 was used. Thirdly, when no further improvement to the fit was possible, backward elimination was carried out to find the smallest model that was as good as the full model. Here a $p$-value of 0.10 was used for the stopping criterion.

### Model development with multiple imputation

In the setting of the multiple imputation process, we suggested two possible strategies that can be applied to carry out the model development process.

#### *Strategy 1*

All three stages of the model development process - the selection of main effects, identification of interactions as well as the backward elimination - are performed on the initial data set generated by the EM parameters. The variables and interactions identified by this process are incorporated into the imputation model. Interactions are treated differently, depending on which imputation method is used.

For MVNI as implemented in the NORM software, interactions with p categories are treated as categorical variables and coded into p-1 dummy variables before being added to the raw incomplete data. By way of an example: an interaction between gender (male/female) and smoking (yes/no) is broken down into separate categories – male/yes, male/no, female/yes and female/no – and binary coding (present/absent) is applied to the first three categories.

For FCS, the interaction is coded according to the possible categories. So, in the example above, male/yes = 1, male/no = 2, female/yes = 3 and female/no = 4.

The interactions as coded in the two scenarios above are merely treated as additional variables. This has been referred to as the 'transform-then-impute' method of dealing with interactions and, in a regression model that includes interactions, has been shown to yield good regression estimates, even though the imputed values are inconsistent with one another. In contrast to this is the 'impute-then-transform' method, also known as passive imputation, which yields plausible-looking imputed values but biased regression estimates [25].

This imputation model is then used to produce the m sets of imputed data. These are analysed individually and the results are combined using Rubin's rules [4].

#### *Strategy 2*

Using the initial EM generated data set, the first two stages of the model development process are completed - selection of main effects and identification of interactions. These are then incorporated into the imputation model as before and m sets of imputed data are produced. Analysis, followed by the third stage of model development (backward elimination), is then applied to each of these data sets. The final selection of variables for the model includes those that are selected in at least 50% of the individual data sets. In the event that no variables satisfy the selection criterion, the condition can be relaxed to a lower percentage. Once these variables are established, analysis is carried out on each data set and the results are combined.

### Analysis

Analyses were carried out using the Statistical Package for Social Sciences (SPSS v17). Given that the outcome variable, asthma severity, is an ordinal measure, the chosen method of analysis for this data was ordinal regression. The three categories of the outcome variable are 'none/mild intermittent asthma'; 'mild persistent asthma' and 'moderate/severe asthma'. For all the analyses, logit was the chosen link function.

In addition to the analysis of the imputed data, a complete case analysis was carried out for comparative purposes. All main effects and interactions that were defined in stages 1 and 2 of the model building process were used with the complete case analysis and then backward elimination was applied to reduce the model.

## Results

### Data review

A total of 22 variables make up the data for this analysis. (1 interval and 21 categorical environmental, genetic and socio-economic variables) (Table 2). Of these variables, 18 (81.8%) experienced some missing data; a total of 166 (43.5%) of the subjects had incomplete data; and, overall, 445 (5.3%) items of data were missing. Missingness in variables ranged from 19.4% to less than 5%. Completely measured variables include age, gender, area and the outcome variable, asthma severity. The missing values follow a nonmonotonic pattern. The majority of non-response was as a result of whole sections or pages of questionnaires being left out. In some instances, one or more of the four questionnaires were missing. There were also numerous cases of seemingly random omissions of individual data items and, in some cases, it is evident that the required information was not known.

Results from the chi-square analysis, to test whether either the incomplete variable or its missingness was related to observed values of other variables, showed that for all but three of the incomplete variables, missingness was associated with measured values in other variables; and all variables were associated with at least one other variable in the set. Thus missingness for these variables can be assumed to be MAR. However, it cannot be ruled out that there exists some MNAR mechanism in the data. Further analysis showed that the distribution of the outcome variable, asthma severity, is the same (in a

**Table 2 Variables, categories and the percentage missing**

| Variable | Response category | % missing |
|---|---|---|
| Gender | male/female | 0 |
| Neonatal care | yes/no | 3.7 |
| Birth weight | up to 2.5 kg/>2.5 kg/don't know | 1.0 |
| Fear in neighbourhood | yes/no | 6.5 |
| Smoked while pregnant | yes/no | 50. |
| Smokers in the home | yes/no | 0.3 |
| Smoke exposure in vehicles | yes/no | 7.6 |
| Exercise | Up to once a week/2-4 times a week/>4 times a week | 6.3 |
| TV watching | Up to an hour a day/1-3 hours a day/>3 hours a day | 6.5 |
| Number people in home | 1-4/5-7/8+ | 9.2 |
| Income (monthly) | up to R1000/R1001-R4500/R4501-R10000/R10001+ | 19.4 |
| Food availability | not always enough/enough | 8.4 |
| Perceived weight | overweight/underweight/correct weight | 6.8 |
| Work and wear | yes/no | 3.7 |
| Pets at home ever | yes/no | 1.0 |
| Area | South Durban/North Durban | 0 |
| Breakfast habits | Not every day/daily | 6.5 |
| Violence experienced | yes/no | 7.3 |
| Attacked with weapons | yes/no | 7.3 |
| Stove type | paraffin/gas/electric/none | 9.9 |
| Age | | 0 |
| Asthma severity | Moderate-severe/mild persistent/mild intermittent/no asthma | 0 |

statistical sense) for whether data is present or missing for all variables except 'food availability', where fewer than expected of those with missing data on the food variable did not have asthma. Because asthma severity is related to the missingness of 'food availability' but not to 'food availability' itself, the inclusion of asthma severity in the imputation model will make the MAR assumption for 'food availability' more plausible [9].

## Model development

### *Imputed data -MVNI*

The two different strategies suggested for building the model using the imputed data resulted in the identical set of variables and interactions being identified. In each case 17 main effects and 10 interactions were included in the final model (Table 3). While fewer than half of the main effects were significant, the interactions in which these variables were involved were largely significant. Main effects dropped from the model include birth weight, perceived weight, weapons and stove type. However, these were left in the imputation model as they were shown to be associated with other variables and/or their missingness.

### *Imputed data -FCS*

Model development following strategy 1 resulted in the identical model as identified when applying MVNI imputation. The set of significant variables from the two analyses were, however, not the same. Two main effects and three interactions differed in their significance. With strategy 2, the variable 'Smoke while pregnant' and its interaction with 'area' did not make the cut to be included in the model. These two variables were significant in only 9 of the 20 individual analyses, whereas, they were significant in 10 of the 20 analyses when MVNI imputation was applied.

### *Complete case analysis*

The complete case analysis was based on 216 complete cases, representing 56.5% of the total available cases. The final model contained 16 main effects and 7 interactions (Table 3).

The main effects selected with the complete case data compared to those selected with the imputed data differed slightly. 'Perceived weight' and 'weapons' are the only variables that are in the complete case model but not in the imputed data model. Three of the 10 interactions and three of the main effects from the imputed data models were not retained in the complete case model. The models from the imputed data contained more variables than the complete case model.

## Analysis

Results of the three different analyses of the imputed data (Table 3) are, in general, very similar. The size and direction of association between asthma severity and all the predictor variables, as well as the standard errors (SE's) of the estimated coefficients are consistent across both types of imputation as well as for both model building strategies. Even though some differences in the significance of certain predictors did occur, in all cases the p-values showing significance of these predictors were only marginally different from the 5% cut-off value.

A comparison of results of the complete case analysis (CC) with the other analyses shows that the standard errors of the estimated coefficients for the CC analysis are appreciably larger in all but the one predictor variable – 'smoke in vehicle'. There are also noticeable differences in the magnitude of the estimated coefficients for the

**Table 3 Estimated coefficients (EST) and standard errors (SE) for the predictors selected in the different analyses**

| Predictor | Reference Category | Category | CC (N = 216) | | MVNI (N = 382) | | FCS1 (N = 382) | | FCS2 (N = 382) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | EST | SE | EST | SE | EST | SE | EST | SE |
| Gender | Female | Male | -0.441 | 0.674 | 0.129 | 0.398 | 0.030 | 0.391 | 0.017 | 0.390 |
| Neonatal care | No | Yes | 2.484* | 0.723 | 1.103* | 0.444 | 1.112* | 0.450 | 1.085* | 0.446 |
| Fear | No | Yes | -1.169 | 0.649 | -0.958* | 0.431 | -1.009* | 0.451 | -1.073* | 0.444 |
| Smoked while pregnant | No | Yes | 4.256* | 1.237 | 1.019 | 0.736 | 0.885 | 0.693 | 0 | |
| Smokers in home | No | Yes | 0.939 | 0.537 | 0.742* | 0.352 | 0.761* | 0.341 | 0.801* | 0.335 |
| Smoke in vehicles | No | Yes | -2.584* | 0.921 | -0.253 | 1.068 | -0.308 | 1.011 | -0.323 | 1.015 |
| Exercise | >4 times a week | Up to once a week | 2.805* | 1.227 | 0.892 | 0.761 | 0.692 | 0.756 | 0.624 | 0.731 |
| | | 2 – 4 times a week | 3.313* | 1.229 | 1.039 | 0.717 | 0.936 | 0.718 | 0.738 | 0.680 |
| TV watching | >3 hours a day | Up to 1 hour a day | -0.566 | 0.854 | 0.399 | 0.684 | 0.327 | 0.669 | 0.346 | 0.657 |
| | | 1 – 3 hours a day | 0.304 | 0.769 | 0.641 | 0.639 | 0.525 | 0.630 | 0.569 | 0.618 |
| Number people in home | 8+ | 1 - 4 | 0 | | 1.084 | 0.554 | 1.060* | 0.539 | 1.101* | 0.526 |
| | | 5 - 7 | 0 | | 0.226 | 0.552 | 0.254 | 0.551 | 0.250 | 0.540 |
| Income | R100001+ | up to R1000 | 2.840* | 1.257 | 0.695 | 0.8 | 0.787 | 0.789 | 0.823 | 0.778 |
| | | R1001 – R4500 | 1.285 | 1.203 | 0.209 | 0.797 | 0.489 | 0.754 | 0.431 | 0.754 |
| | | R4501 – R10000 | 1.933 | 1.17 | 1.428 | 0.783 | 1.401* | 0.692 | 1.356 | 0.692 |
| Food availability | Enough | Not always enough | -0.575 | 0.64 | 0.604 | 0.503 | 0.665 | 0.464 | 0.677 | 0.455 |
| Perceived weight | Correct weight | Overweight | -0.230 | 0.743 | 0 | | 0 | | 0 | |
| | | Underweight | 2.369* | 0.97 | 0 | | 0 | | 0 | |
| Work'nWear | No | Yes | 0 | | -0.635 | 0.626 | -0.543 | 0.629 | -0.478 | 0.622 |
| Pets ever | No | Yes | -3.770* | 0.994 | 1.658* | 0.501 | -1.483* | 0.503 | -1.413* | 0.467 |
| Area | North Durban | South Durban | 6.278* | 1.461 | 2.042* | 0.76 | 1.948* | 0.737 | 1.597* | 0.671 |
| Breakfast habits | Daily | Not daily | -4.098 | 3.04 | -0.492 | 1.512 | -0.234 | 1.548 | -0.110 | 1.518 |
| Violence | No | Yes | 0 | | -0.817* | 0.382 | -0.741* | 0.377 | -0.715 | 0.373 |
| Weapons | No | Yes | -1.147* | 0.555 | 0 | | 0 | | 0 | |
| Age | | | -1.068* | 0.438 | -0.79* | 0.254 | -0.833* | 0.268 | -0.834* | 0.265 |

**Table 3 Estimated coefficients (EST) and standard errors (SE) for the predictors selected in the different analyses** *(Continued)*

| Predictor | Reference Category | Category | CC (N = 216) | | MVNI (N = 382) | | FCS1 (N = 382) | | FCS2 (N = 382) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | EST | SE | EST | SE | EST | SE | EST | SE |
| Fear*Breakfast | No/daily | Yes/not daily | 2.635* | 1.219 | 2.047* | 0.866 | 2.123* | 0.916 | 2.185* | 0.911 |
| Gender*SmokeVehicle | Female/No | Male/yes | 5.092* | 1.342 | 2.535* | 1.034 | 2.431* | 0.977 | 2.464* | 0.971 |
| SmokeVehicle*TV | No/>3 hrs | Yes/up to 1 hr | 0 | | 0.891 | 1.298 | 0.675 | 1.265 | 0.722 | 1.250 |
| | | Yes/1 – 3 hrs | 0 | | -2.184* | 1.085 | -1.975 | 1.034 | -2.002 | 1.037 |
| Food*Age | enough/ | Not always enough/ | 1.762* | 0.743 | 0.925* | 0.396 | 0.786* | 0.385 | 0.778* | 0.364 |
| Exercise*Area | >4 times/ND | < once a week/SD | -4.573* | 1.533 | -1.41 | 1.031 | -1.255 | 0.954 | -1.125 | 0.923 |
| | | 2 – 4 times/SD | -6.331* | 1.627 | -1.981* | 0.913 | -1.805* | 0.896 | -1.551 | 0.850 |
| Income*Breakfast | > R10000/daily | ≤R1000/not daily | -4.051 | 2.5 | -3.921* | 1.8 | -3.666* | 1.731 | -3.808* | 1.733 |
| | | R1001-R4500/not daily | 0.414 | 2.408 | -1.218 | 1.636 | -1.439 | 1.530 | -1.513 | 1.516 |
| | | R4501-R10000/not daily | 2.479 | 2.395 | -1.374 | 1.541 | -1.568 | 1.454 | -1.715 | 1.431 |
| TV*Breakfast | >3 hrs/daily | ≤1 hr/not daily | 6.310* | 2.213 | 2.573* | 1.259 | 2.051 | 1.192 | 1.976 | 1.186 |
| | | 1-3 hrs/not daily | 1.974 | 2.154 | 0.192 | 1.109 | 0.270 | 1.112 | 0.192 | 1.103 |
| SmokeVehicle*Age | no/ | yes/ | 0 | | 0.814* | 0.375 | 0.809* | 0.348 | 0.782* | 0.341 |
| Smoke preg*Area | no/ND | yes/SD | -5.118* | 2.101 | -1.875 | 1.363 | -1.663 | 1.291 | 0 | |
| Work'nWear*Breakfast | no/not daily | yes/daily | 0 | | 2.349* | 1.076 | 2.095 | 1.070 | 2.165* | 1.090 |

ND – North Durban; SD – South Durban; preg – pregnant.
CC – Complete case.
MVNI – Multiple imputed MVNI strategies 1 and 2.
FCS1 -Multiple imputed FCS strategy 1.
FCS2 -Multiple imputed FCS strategy 2.
*Significant at the 0.05 level.

CC analysis as compared to the other analyses. Contradictions are also present regarding the relationship with asthma severity for some of the predictors.

### Diagnostics

In order to confirm that the imputed values are reasonable, each variable with missing data in excess of 8% was examined to identify variables with large differences between the measured and imputed. The variables considered included income, stove type, number of people and food availability (Figure 1). The Kolmogorov-Smirnov test was applied to assess whether significant differences exist between the distributions of the imputed data – both MVNI imputed and FCS imputed – and the measured data [26]. No significant differences were found.

In analysis testing for significant differences between the distributions of the imputed data sets and the complete case data, no significant differences were found.

Another useful diagnostic that gives an indication of the stability of the estimates resulting from multiple imputation is the degrees of freedom (df) associated with the t-value in Rubin's rules and adapted from Schafer [8,9]. The df associated with multiple imputation is not the same as the df found in other statistical concepts and rather is a 'measure' of the ratio of the within-imputation variance to the between-imputation variance. In this study, df ranged from 130.54 to 9073.51 for the NORM imputations and from 138.88 to 15135.431 for the FCS imputations which, being large compared to the number of imputed sets, is an indication that the estimates have stabilized and can be trusted.

## Discussion

In this study investigating methods for addressing missing data, specifically when including interactions in the analysis, we found support for building the model using an EM generated set of data and then applying multiple imputation as a robust method to address this common shortcoming in epidemiological studies.

Epidemiological studies frequently suffer from missing data. Many researchers avoid this problem by dropping all cases with data missing on any variable and carrying out what is known as a complete case analysis. An advantage of this type of analysis is that it is computationally easy to apply and can be done with any reputable commercial software package. However, unless the data is MCAR, the values of the estimated coefficients produced with this analysis may be biased. Moreover, when the missingness is not only a function of the covariate(s) but also of the outcome variable, then the bias from a complete case analysis is heightened [27]. Although complete case analysis and other *ad hoc* methods, like mean substitution and the missing-indicator method, are still widely used, researchers are becoming more aware of the perils of applying such methods and many are now employing multiple imputation methods to address

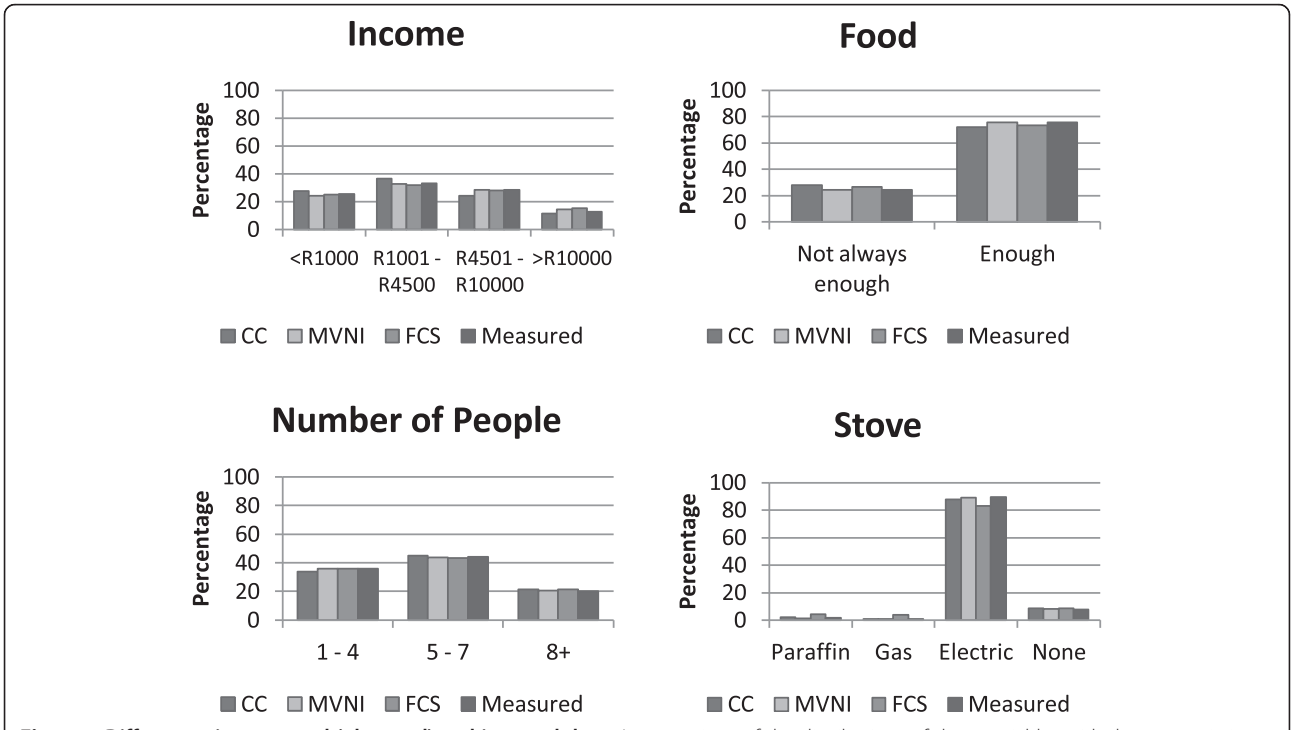



**Figure 1 Differences in measured (observed) and imputed data.** A comparison of the distributions of the 4 variables with the most missing data for the complete case data (CC), MVNI imputed data, FCS imputed data and measured data.

the missingness in their data. While results from multiple imputation will be unbiased when data is MAR, it has been suggested that even when it is MNAR, adequately dealing with as much of the missingness mechanism as possible will usually produce sound results [8-11]. This is achieved by including auxiliary variables – those variables related to the missingness but not necessarily included in the analysis, interactions and the outcome variable in the imputation model.

While much has been published on the application of multiple imputation to epidemiological studies, there is limited literature that deals with model building in the presence of missing data, and more specifically model building including interactions. The aim of this paper was to demonstrate a simple and easily applied strategy to build interactions, which are not known up front, into a model while at the same time imputing the missing data.

The dilemma that we faced was a practical one. It is possible for the interactions to be added after imputation. This is termed passive imputation or 'impute-then-transform'. However, it has been shown that including interactions, as product terms, before imputation produces superior results than if the imputations are done first and the interactions are added at the analysis stage [25]. For the best results, the identified interactions should be included in the imputation model along with the predictor variables, the auxiliary variables and the outcome variable. However, how can the interactions be identified and the best model built, when the data is incomplete?

Two strategies for model building, S1 and S2, were explored – both utilizing a single imputed data set generated from the ML parameter estimates produced from the EM algorithm for covariance matrices.

Imputation was carried out with both multivariate normal imputation (MVNI) and the more flexible fully conditioned specification (FCS). The same set of 17 predictor variables and 10 interactions for the best model were identified when applying MVNI with both strategies S1 and S2, as well as with the application of FCS and strategy S1. FCS with strategy S2 failed to include one of these predictors and an associated interaction in its best model. Since these dropped variables did not alter the interpretation of the results, it would seem that both strategies for model building are equally effective. The advantage of S1 over S2 is that it is easier and less time-consuming to execute and therefore probably the preferred choice.

In comparison to the model variables selected from the imputed data, fewer variables were selected for the model on the complete case data. This is most likely caused by the enormous reduction in cases and the subsequent loss of power.

A total of 5.3% missing items spread across 81.8% of variables, affecting 43.5% of cases was present in the dataset used for this analysis. Examination of the missingness revealed that it is possible that the missingness mechanism present in this data is a combination of MCAR, MAR and MNAR. Analysis of the relationships between both the missingness of the variables and the variables themselves confirmed that significant relationships exist between each of the variables and at least one other variable in the set; furthermore, the missingness of all but three of the variables is significantly related to at least one other variable in the set.

For reliable and unbiased results to be obtained from a complete case analysis, the data is required to be MCAR, which is clearly not the case here. Furthermore, although this means of dealing with missing data is acceptable when the lost cases amount to no more that 5%, this data set is reduced by over 40% which will inevitably have a negative effect on the outcome of the analysis.

On the other hand, multiple imputation, if applied correctly, is able to produce sound results when the data is MAR and it has been shown that even when the data is MNAR, the effects of this mechanism are often surprisingly minimal [11]. In order to ensure that the imputation model was general enough to encompass the subsequent analysis, the outcome variable, interactions and variables related to either the incomplete variables or their missingness or both were included in the imputation model. By including variables that are correlated with each incomplete variable but not its missingness, we expect that the additional information will cause a decrease in the standard errors and hence an increase in efficiency and statistical power [10]. If there is an element of MNAR present in the data, the inclusion of these variables in the imputation model should lessen the bias and make the assumption of MAR more plausible.

It is unclear as to how many variables and interactions, given the sample size available, can be reliably assessed with multiple imputation applications. It seems that this depends to some extent on the software being used. In some cases, convergence of large models is a problem in that it can make the imputation process unacceptably slow [16]. Graham and Schafer [28], in a study using NORM to perform the imputations found that results were quite acceptable "even with sample sizes as low as 50, even with as much as 50% missing from most variables, and even with relatively large and complex models". In a study on the imputation of categorical data [29] it was found that, while problems exist when imputing using a variant of NORM designed to deal with categorical data when many variables are present, the same limitations are not problematic for NORM. In another study [30] on the inclusion of continuous auxiliary

variables in the imputation model, the authors suggest the ratio of cases with complete data to variables should be at least 3:1. Given these guidelines, we found that convergence for both imputation methods was achieved quickly and reliably. Furthermore, even with the dummy coding of all the categorical variables and the interactions, the ratio of complete cases to variables far exceeds 3:1. We are therefore confident that our results are reliable.

Diagnostic tests on the distributions of the imputed data showed that data imputed both with MVNI and FCS were not significantly different from either the measured data or the CC data. These results confirm findings that multiple imputation with MVNI incorporating sensible rounding should work in most situations [14], even in the presence of binary and ordinal variables [6].

The diagnostic measure, df, also indicated that the estimates obtained from both multiple imputation methods have stabilized and are therefore trustworthy.

Analysis of the two sets of imputed data yielded very similar results. This is consistent with findings from a study comparing the two imputation approaches [6] where it was found that "similar results can be expected from FCS and MVNI in a standard regression analysis involving variously scaled variables". The magnitude of the standard errors and the magnitude and direction of the estimated coefficients were consistent across both these imputation types and for both model building strategies. While there were some inconsistencies in the significance of predictors, these did not affect the overall interpretation of the associations between asthma severity and the factors included on the models.

A comparison of results for the complete case analysis and the analyses of the imputed data showed that standard errors for the estimated coefficients from the analysis of the imputed data were, in all but one case, considerably smaller than those from the complete case analysis. These smaller standard errors resulted in greater accuracy of the estimated coefficients. This increased precision indicates the superior efficiency and statistical power obtained for the analysis of the imputed data. The inconsistencies in the signs of the estimates and the significance of the predictors could result from the non-random fashion in which cases are dropped for the complete case analysis which may distort the joint distribution among the variables. The resulting bias in point estimates could lead to misidentification of significant predictors [31]. Another important factor that would negatively affect results of the complete case analysis is that the missingness mechanism present in the data is not confined to being MCAR. While multiple imputation methods produce unbiased parameter estimates when the missingness is MAR, this is not the case with complete case analysis. This missingness mechanism factor could also have added to the large difference in magnitude of the standard errors for the complete case analysis as compared to the imputed data analysis that, some would argue, could not be explained on the basis of sample size alone.

These results are consistent with what we expect given the significant reduction in cases for the complete case analysis and the missingness mechanism present in the data that would almost certainly result in a loss of power and the introduction of bias into estimates.

Given the rigid processes followed in the imputation of the data and subsequent analyses, we would suggest that the results from the imputed data can be considered reliable. On the other hand, the results from the complete case analysis should be treated with caution.

## Conclusions

With the development of readily available and easily implemented software, multiple imputation methods for dealing with missing data are becoming more popular in epidemiological studies that have incomplete measured variables. A critical part of the imputation process is the inclusion of those variables that are correlated with missingness as well as the interactions to be used in the analysis process. While this can present a practical challenge if the interactions are not specified *a priori*, we have illustrated one possible approach that effectively identifies the best main effects and interactions for a model in the presence of missing data and at the same time, imputes the data items that are missing. Undoubtedly, further testing of these strategies on other data sets is needed. It is hoped that the ideas presented in this paper can be further explored and developed so that, by addressing this practical dilemma, more medical researchers will be able to apply multiple imputation when data suffers from missingness.

#### Abbreviations

CC: Complete case; df: Degrees of freedom; EM: Expectation Maximization; EST: Estimates; FCS: Fully conditional specification; MAR: Missing at random; MCAR: Missing completely at random; MI: Multiple imputation; MNAR: Not missing at random; MVNI: Multivariate normal imputation; ND: North Durban; SD: South Durban; SE: Standard errors.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

GH analysed and interpreted the data, drafted the original article and made suggested revisions. RN was involved in the collection of the data, supervised the study, critically reviewed the article and approved the final version to be published. GM was involved in the collection of the data and critically reviewed the article. DN and TZ supervised the study and reviewed the original article. All authors read and approved the final manuscript.

#### Acknowledgments

This work was supported by eThekwini Metropolitan Municipality (local government) – Contract No 1A-103; Medical Research Council of South Africa and University of KwaZulu-Natal – Research Funds. The authors are grateful

to the editor and reviewers for their helpful comments and suggestions which resulted in numerous improvements.

**Author details**
[1]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville Campus, University Road, Westville, Durban, South Africa. [2]Discipline of Occupational and Environmental Health, School of Nursing and Public Health, University of KwaZulu-Natal, Durban, South Africa. [3]Department of Environmental Health Sciences, School of Public Health, University of Michigan, 6655 SPH I, Ann Arbor, MI 48109-2029, USA.

Received: 1 August 2014 Accepted: 26 November 2014
Published: 19 December 2014

**References**

1. Klebanoff MA, Cole SR: **Use of multiple imputation in the epidemiologic literature.** *Am J Epidemiol* 2008, **168**(4):355–357.
2. Greenland S, Finkle WD: **A critical look at methods for handling missing covariates in epidemiologic regression analyses.** *Am J Epidemiol* 1995, **142**(12):1255–1264.
3. Little RJA, Rubin DB: *Statistical Analysis With Missing Data.* New York: J. Wiley; 1987.
4. Rubin DB: *Multiple imputation for nonresponse in surveys.* New York: Wiley; 1987.
5. Graham JW: **Missing data analysis: making it work in the real world.** *Annu Rev Psychol* 2009, **60:**549–576.
6. Lee KJ, Carlin JB: **Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation.** *Am J Epidemiol* 2010, **171**(5):624–632.
7. Donders ART, van der Heijden GJ, Stijnen T, Moons KG: **Review: a gentle introduction to imputation of missing values.** *J Clin Epidemiol* 2006, **59**(10):1087–1091.
8. Schafer JL, Olsen MK: **Multiple imputation for multivariate missing-data problems: a data analyst's perspective.** *Multivariate Behav Res* 1998, **33**(4):545–571.
9. Graham JW: **Missing data: Analysis and Design.** New York: Springer; 2012.
10. Collins LM, Schafer JL, Kam C-M: **A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures.** *Psychological Methods* 2001, **6:**330–351.
11. Graham JW, Hofer SM, Donaldson SI, MacKinnon DP, Schafer JL: **Analysis with missing data in prevention research.** In *The science of prevention: methodological advances from alcohol and substance abuse research.* Washington D.C.: American Psychological Association; 1997:325–366.
12. Rubin DB: **Multiple imputation after 18+ years.** *J Am Stat Assoc* 1996, **91**(434):473–489.
13. Stuart EA, Azur M, Frangakis C, Leaf P: **Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative.** *Am J Epidemiol* 2009, **169**(9):1133–1139.
14. Schafer J: *Analysis of incomplete multivariate data.* London: Chapman & Hall; 1997.
15. Vergouwe Y, Royston P, Moons KG, Altman DG: **Development and validation of a prediction model with missing predictor data: a practical approach.** *J Clin Epidemiol* 2010, **63**(2):205–214.
16. White IR, Royston P, Wood AM: **Multiple imputation using chained equations: issues and guidance for practice.** *Stat Med* 2011, **30**(4):377–399.
17. Wood AM, White IR, Royston P: **How should variable selection be performed with multiply imputed data?** *Stat Med* 2008, **27**(17):3227–3246.
18. Naidoo RN, Robins TG, Batterman S, Mentz G, Jack C: **Ambient pollution and respiratory outcomes among schoolchildren in Durban, South Africa.** *SAJCH* 2013, **7**(4):127–134.
19. Schafer J: *NORM: Multiple imputation of incomplete multivariate data under a normal model [Computer software].* University Park: Pennsylvania State University, Department of Statistics; 1999.
20. Allison PD: **Missing data.** Thousand Oaks, CA: SAGE; 2002.
21. SPSS inc: **Build Better Models When You Fill in the Blanks.** 2014. Available from: http://www.spss.com/media/collateral/statistics/missing-values.pdfs.
22. Azur MJ, Stuart EA, Frangakis C, Leaf PJ: **Multiple imputation by chained equations: what is it and how does it work?** *Int J Methods Psychiatr Res* 2011, **20**(1):40–49.
23. Raghunathan TE, Solenberger PW, Van Hoewyk J: *IVEware: Imputation and variance estimation software.* Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan; 2002.
24. Graham JW, Olchowski AE, Gilreath TD: **How many imputations are really needed? Some practical clarifications of multiple imputation theory.** *Prev Sci* 2007, **8**(3):206–213.
25. Von Hippel PT: **How to impute interactions, squares, and other transformed variables.** *Sociol Methodol* 2009, **39**(1):265–291.
26. Abayomi K, Gelman A, Levy M: **Diagnostics for multivariate imputations.** *J R Stat Soc Ser C Appl Stat* 2008, **57**(3):273–291.
27. Desai M, Esserman DA, Gammon MD, Terry MB: **The use of complete-case and multiple imputation-based analyses in molecular epidemiology studies that assess interaction effects.** *Epidemiol Perspect Innovat* 2011, **8**(1):5.
28. Graham JW, Schafer JL: **On the performance of multiple imputation for multivariate data with small sample size.** *Statistical strategies for small sample research* 1999, **50:**1–27.
29. Finch WH: **Imputation methods for missing categorical questionnaire data: a comparison of approaches.** *J Data Sci* 2010, **8**(3):361–378.
30. Hardt J, Herke M, Leonhart R: **Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research.** *BMC Medical Research Methodology* 2012, **12**(1):184.
31. He Y: **Missing data analysis using multiple imputation getting to the heart of the matter.** *Circ Cardiovasc Qual Outcomes* 2010, **3**(1):98–105.

doi:10.1186/1471-2288-14-136
**Cite this article as:** Hendry *et al.*: Model development including interactions with multiple imputed data. *BMC Medical Research Methodology* 2014 **14**:136.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

# The Effect of the Mechanism and Amount of Missingness on Subset Correspondence Analysis

Gillian M. Hendry*[1], Temesgen Zewotir[1], Rajen N. Naidoo[2], Delia North[1]

[1] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville Campus, University Road, Westville, Durban, South Africa

[2] Discipline of Occupational and Environmental Health, School of Nursing and Public Health, University of KwaZulu-Natal, Durban, South Africa

Email addresses:

GMH: hendryfam@telkomsa.net

TZ: Zewotir@ukzn.ac.za

RNN: naidoon@ukzn.ac.za

DN: northd@ukzn.ac.za

* Corresponding author

Gillian M Hendry

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville Campus, University Road, Westville, Durban, South Africa
hendryfam@telkomsa.net

**Short title: Effect of missingness mechanism on Subset CA**

## Abstract

The application of subset correspondence analysis is a relatively new technique to deal with the analysis of categorical data that suffers from missingness. This simulation study tests the effects of Little and Rubin's missingness mechanisms, as well as missingness up to 50% on the analysis of data using sCA. Missingness was simulated across 18 different scenarios and each scenario was repeated 10 times, with outcomes averaged across the 10 simulations. It was found that while missingness in excess of 30% has some effect on certain outcomes, there is no evidence to suggest that the missingness mechanism significantly affects results.

## Introduction

Missing data is common in many studies and presents a challenge, in particular if the data is categorical in nature. Missingness is traditionally categorized as missing completely at random (MCAR), in which each data item has an equal chance of being missing; missing at random (MAR), in which the missingness in a variable is dependent on another known and measured variable; and missing not at random (MNAR), in which the missingness in a variable depends on the value of the data item itself (Little & Rubin, 1987). Many *ad hoc* approaches are used to address the issue of missingness but these are, on the whole, not recommended as results may be biased unless data is MCAR.

More recently, multiple imputation (MI) has become a recommended method to deal with missing data and will produce unbiased estimates so long as the data is MAR (Donders, van der Heijden, Stijnen, & Moons, 2006). There are conflicting opinions regarding the treatment of data that is MNAR. It has been suggested that when data is MNAR, "there is no universal method of handling the missing data properly" (Donders et al., 2006; Greenland & Finkle, 1995; Rubin, 2004). It is thought that missingness which is not completely at random must be explicitly modelled to obtain unbiased results and this can be a challenging exercise (Gelman & Hill, 2007;

Little & Rubin, 1987; Rubin, 2004). Some have suggested that, with careful application, MI will work with data that is MNAR (Graham, 2009; Van Buuren, 2007). It has been shown that as long as the missingness is kept to under 25%, multiple imputation will produce acceptable results (Scheffer, 2002). Furthermore, a sensitivity analysis carried out by Graham et al (Graham, Hofer, Donaldson, MacKinnon, & Schafer, 1997) found that the effects of an MNAR mechanism is often 'surprisingly minimal' when multiple imputation is applied (Wayman, 2003).

An alternative approach that has been shown to successfully address the issue of missing categorical data is the application of subset correspondence analysis (sCA)(Greenacre & Pardo, 2006; Hendry, North, Zewotir, & Naidoo, 2014).

sCA is the application of correspondence analysis (CA) on a subset of the data – in this case the subset of measured data. A graphical, exploratory technique, sCA reduces a matrix in multi-dimensional space to a subspace of lower dimension, such that the variation (distance between the row and column points) is maximized. In CA and its variants, this variation is termed inertia.

While many studies have been done on the effect of different missingness mechanisms on both *ad hoc* and multiple imputation methods for dealing with missing data (Little & Rubin, 1987; Marshall, Altman, Royston, & Holder, 2010; Peyre, Leplège, & Coste, 2011; Rubin, 2004; Scheffer, 2002; Shrive, Stuart, Quan, & Ghali, 2006; Vach & Blettner, 1991) , we are not aware of any such study with regard to sCA and its use with missing data.

The aim of this simulation study is to explore the effect of both the missingness mechanism and the amount of missingness present when sCA is applied to a set of categorical data that suffers from missingness.

## Methods

The data

A study on the respiratory health of children in Durban, South Africa was undertaken in 2004 (Naidoo, Robins, Batterman, Mentz, & Jack, 2013). The original data, gathered from several schools in the south Durban and north Durban regions, consisted of a number of generic, socio-economic, environmental and behavioural variables as well as a measure of asthma severity. A subset of this data involving 368 cases with complete data across 6 selected variables is used for this study (Table 1).

**Table 1: Categories, code names and frequencies for all variables**

| Variables | Categories | Code names | Count (N = 368) |
|---|---|---|---|
| Age | 8 - 9 years | A1 | 24 |
| | 10 years | A2 | 186 |
| | 11 years | A3 | 134 |
| | 12+ years | A4 | 24 |
| Gender | Male | MAL | 149 |
| | Female | FEM | 219 |
| Neonatal | Yes | NNY | 50 |
| | No | NNN | 318 |
| Smokers | Yes | SY | 180 |
| | No | SN | 188 |
| Area | south Durban | DS | 177 |
| | north Durban | DN | 191 |
| Asthma severity | None/ Mild intermittent | ASNI | 296 |
| | Mild persistent | ASMP | 45 |
| | Moderate/severe | ASMS | 27 |

The purpose of analysis is to explore the relationships between 'age' (categorized into 4 levels from 9 years to 12+ years); 'gender' (M/F); 'neonatal' (whether or not special neonatal care was received at birth); 'smokers' (the presence of smokers in the home) and 'area' (north or south Durban) and their association with 'asthma severity' (none/mild intermittent; mild persistent; moderate to severe).

Missing data mechanisms

To explore the effect of missingness mechanisms (MM) and amount of missingness present (M%), 18 scenarios were considered, with each scenario simulated 10 times. Three MM's were imposed – MCAR, MAR and MNAR - and missingness was generated at rates of 5%, 10%, 20%, 25%, 30% and 50% for each mechanism. Two variables – 'neonatal' and 'smokers' were selected to experience missingness and data was deleted from each of these variables for each scenario.

For the six MCAR scenarios, data was deleted randomly across all categories for each of the variables.

To simulate the MAR mechanism, missingness was imposed on the 'neonatal' and 'smoking' variables according to their association with 'area' and 'gender' respectively. Data was randomly deleted from the 'neonatal' variable such that 30% came from north Durban and 70% from south Durban. Random deletion on the 'smoker' variable was in the ratio 30:70 for M:F. These deletions were completed for each of the six amounts of missingness.

The MNAR mechanism was simulated so that the missing data depended on the actual value of the data item. Deletion from the 'neonatal' variable was carried out such that 10% of required deletions were from the variable category NNY and 90% from variable category NNN. In a similar manner, deletions from the 'smoker' variable involved randomly deleting 90% of required deletions from SY and 10% from SN. Again this was repeated for the six amounts of missingness.

Analysis and outcomes of interest

sCA was applied to each of the simulated data sets and several outcomes were examined to identify effects of the MM and M% on this method. These included:

- COR - relative contributions that the axis makes to the inertia ( variance) of the points
- CTR - absolute contributions that the points make to the inertia of the axis

- TOTINR - a measure of the degree of variation in the measured data
- TI%FULL- the proportion that TOTINR is of the total inertia from an analysis which includes both the measured and the missing data, coded as separate 'missing' categories

Repeated measures ANOVA was applied to the above outcomes to test for significant differences across missingness mechanisms and amount of missingness.

## Results

Full analysis

For the purpose of comparison, sCA was applied to the 368 data set with all variables fully measured. The data is in the form of a contingency table with the three asthma categories as rows and the five selected variables (12 variable categories) as columns.

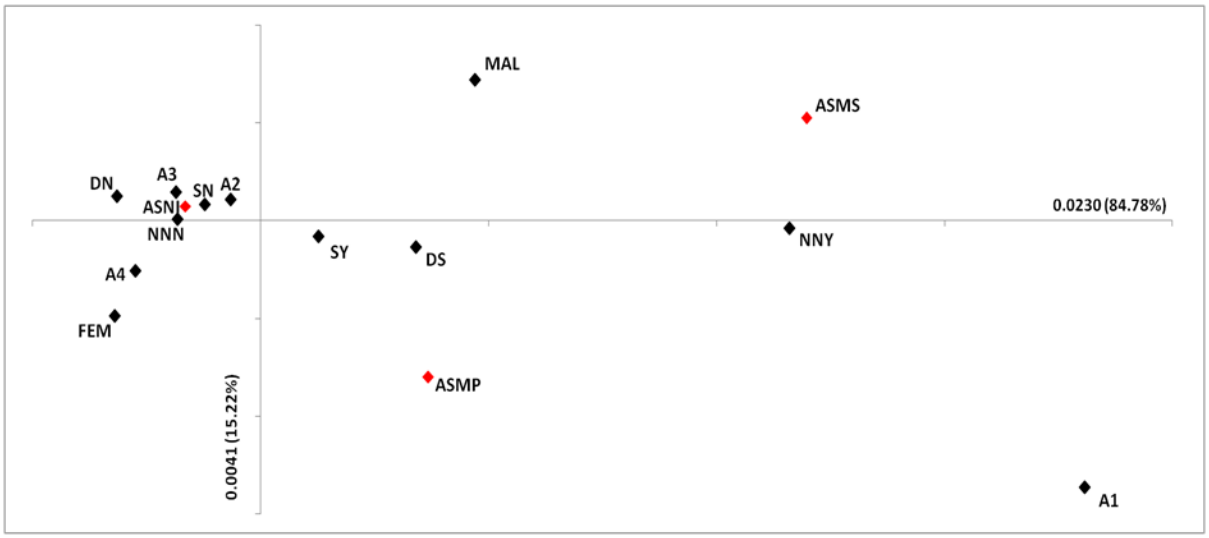



**Figure 1**. Subset correspondence analysis map of the completely measured 368 data set with row points presented by ◆ and column points by ◆. Values on the axes represent principal inertias and their respective percentages of total inertia. Labels as specified in Table 1.

| **Table 2**: **Decomposition of inertia for the two principal axes** | | | | | | |
|---|---|---|---|---|---|---|
| **Name** | **k = 1** | **COR** | **CTR** | **k = 2** | **COR** | **CTR** |
| A1 | 723 | 875 | 297 | -273 | 125 | 236 |
| A2 | -26 | 606 | 3 | 21 | 394 | 10 |
| A3 | -74 | 864 | 17 | 29 | 136 | 15 |
| A4 | -110 | 819 | 7 | -52 | 181 | 8 |
| MAL | 188 | 631 | 125 | 144 | 369 | 407 |
| FEM | -128 | 631 | 85 | -98 | 369 | 277 |
| NNY | 464 | 1000 | 255 | -8 | 0 | 0 |
| NNN | -73 | 1000 | 40 | 1 | 0 | 0 |
| SY | 51 | 910 | 11 | -16 | 90 | 6 |
| SN | -49 | 910 | 11 | 16 | 90 | 6 |
| DS | 136 | 961 | 78 | -27 | 39 | 18 |
| DN | -126 | 961 | 72 | 25 | 39 | 16 |
| ASNI | -66 | 952 | 153 | 15 | 48 | 43 |
| ASMP | 147 | 458 | 116 | -160 | 542 | 762 |
| ASMS | 479 | 954 | 732 | 105 | 46 | 195 |
| K=… coordinates<br>COR relative contributions of inertia<br>CTR absolute contributions of inertia | | | | | | |

Results (Figure 1 and Table 2) show that total inertia across the full subspace of two dimensions is 0.0271, thus indicating that there is limited variability in the data. CTR values, a measure of the absolute contributions of the points to the inertia of the dimension, indicate that variable categories important to the orientation of axis 1 are A1, MAL, NNY and to a lesser extent FEM, DS and DN. This axis separates the lowest asthma severity category (ASNI) on the left from the higher asthma severity categories (ASMP and ASMS) on the right. Associated with the latter categories are A1, MAL, NNY and DS. Variables that play an important part in the orientation of axis 2 are A1, MAL and FEM. This axis separates out ASMS from the other asthma categories, thus enabling a distinction between ASMP and ASMS. Associated with ASMP are FEM and A1. Variables that do not exhibit much variance are situated near the origin. They do not play an important role in the orientation of the axes. These include A2, A3, A4, SY, SN, NNN and ASNI. Associated with the lowest asthma severity classification are A2, A3, SN, DN and NNN.

COR values indicate that axis 1 is more important in terms of contributions to inertia for all variable categories, except ASMP.

Simulated study

*Relative contributions to inertia (COR)*

Average COR values for each missingness mechanism and across the six amounts of missingness are shown for each variable category in Figure 2.

COR values indicate the amount that each axis contributes to the inertia of the point. This makes it possible to identify the axis which contributes most to the inertia of each point. These values are scaled to add to 1000 across all dimensions. Because there are only two possible dimensions for this analysis, and axis 1 accounts for more than 80% of the total inertia, only the COR values for axis 1 are examined. Of the 15 variable categories, only one (ASMP) has a higher COR value on axis 2.

For the fully measured variable categories of 'age', 'gender' and 'area', no significant differences were found in COR values either across MM or for different M%. There are also no significant differences across for the asthma severity categories. However, significant decreases in COR values were found for ASNI and ASMP at 50% missingness.

Examining results for the variables with missingness, while the MM's do not show evidence of significant differences for the smoking category, SY, there are significant differences in the way these mechanisms behave for SN. COR values for MNAR are significantly higher than for the other mechanisms and closer to the 'true' values. With regard to the amount of missingness, when compared to values at 5%, there is a significant reduction in the COR value for MNAR at 50% on SY and from 25% for MCAR on SN.

There are no significant differences across MM or M% for the NNY variable category. While no significant differences were found across MM for the NNN variable category, there is a significant drop in the COR values for MNAR from the 20% missingness point.

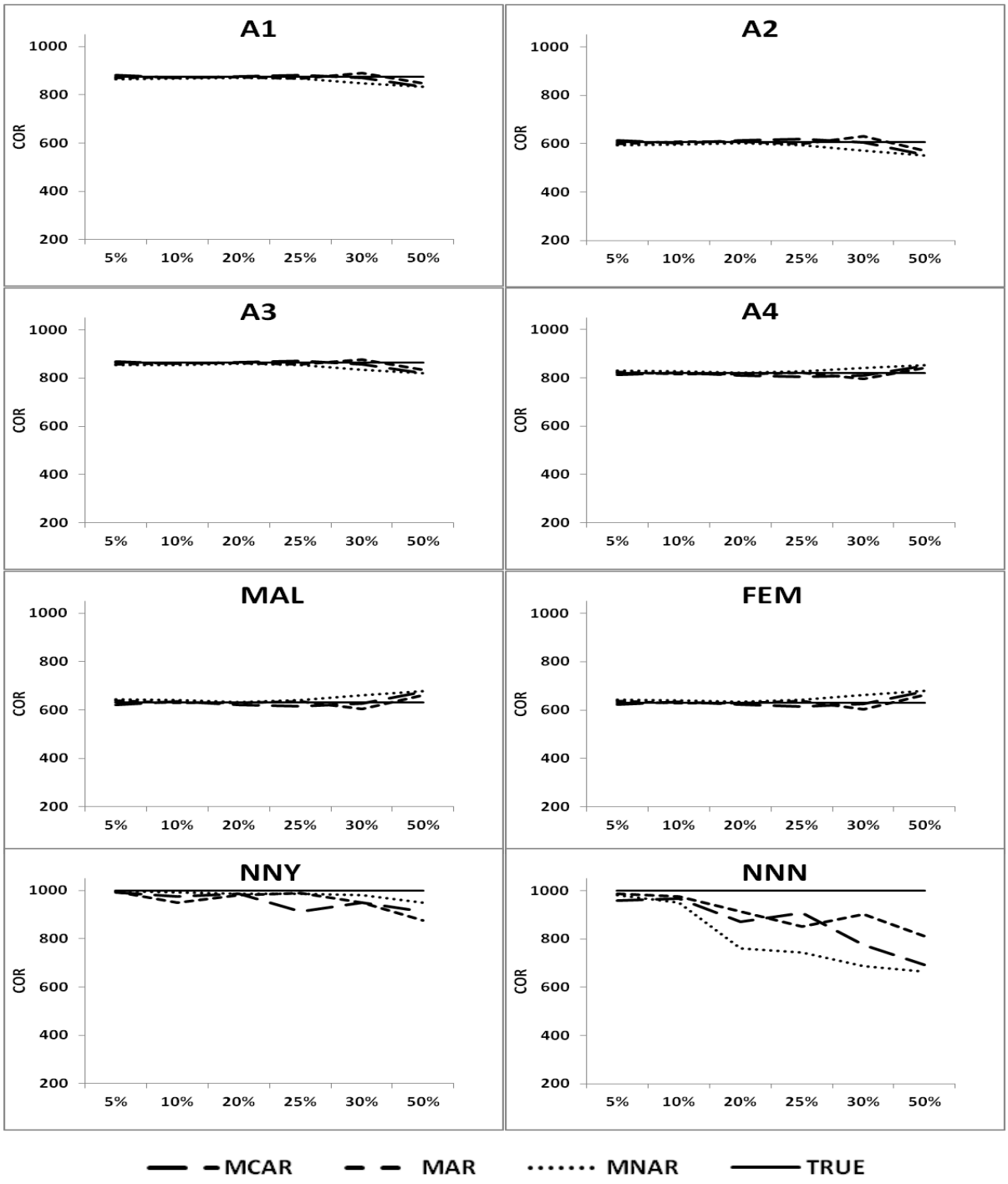



**Figure 2: COR values for each variable across all scenarios**

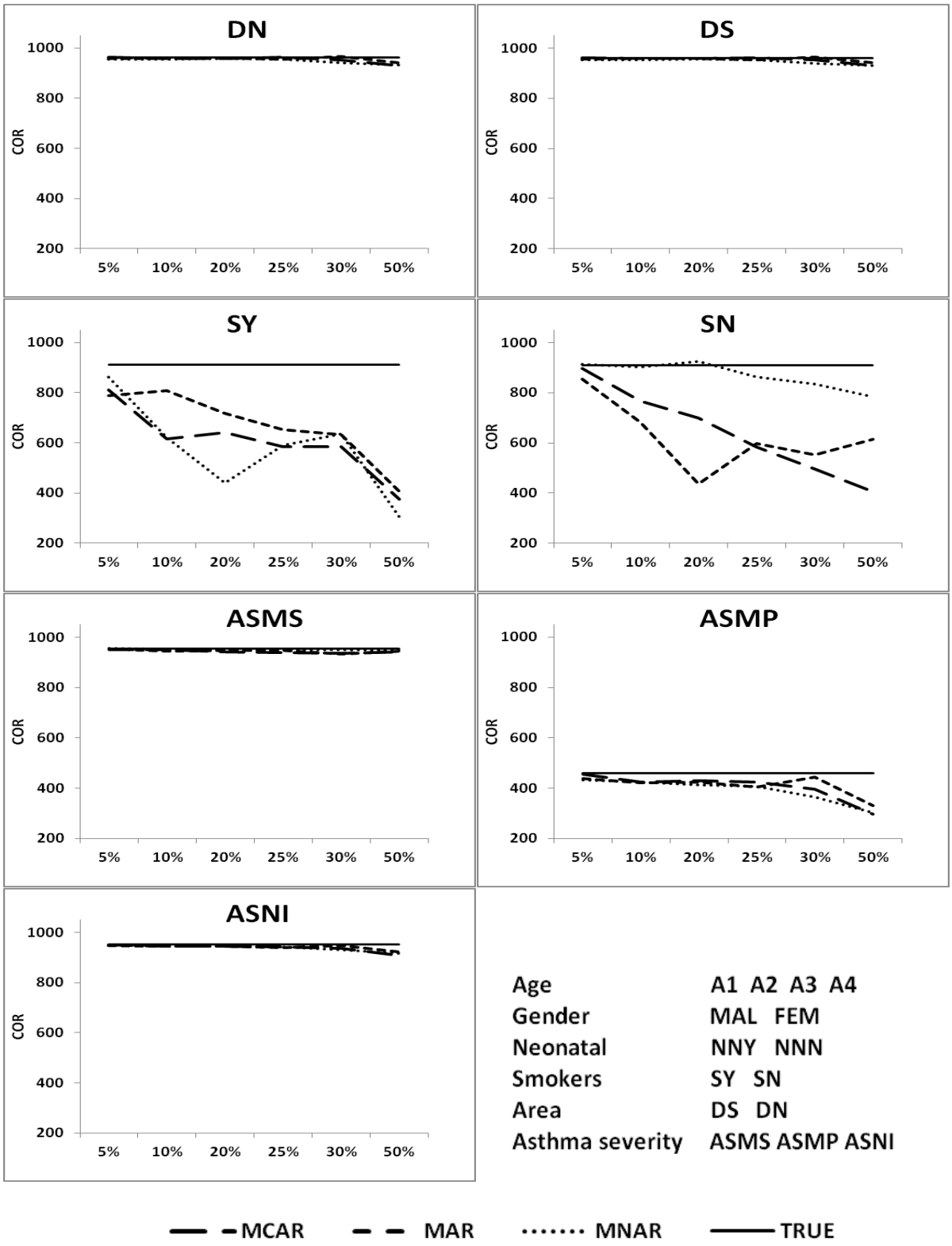



**Figure 2(continued): COR values for each variable across all scenarios**

*Absolute contributions to inertia (CTR)*

Average CTR values for each MM and across the six M% are shown for each variable category in Figures 3 and 4.

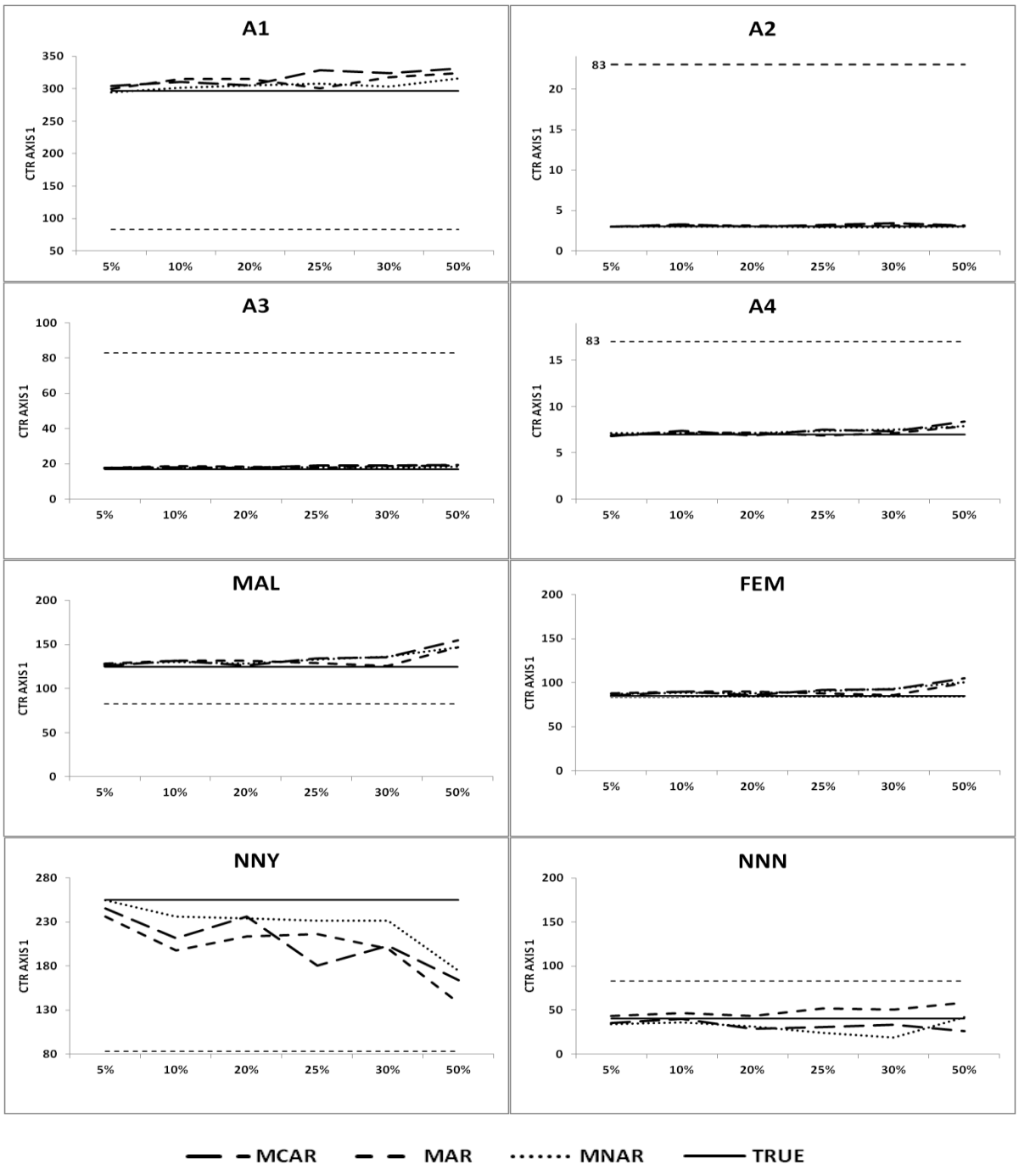



**Figure 3: CTR values for all variables on axis 1 (Dotted line indicates the threshold value)**

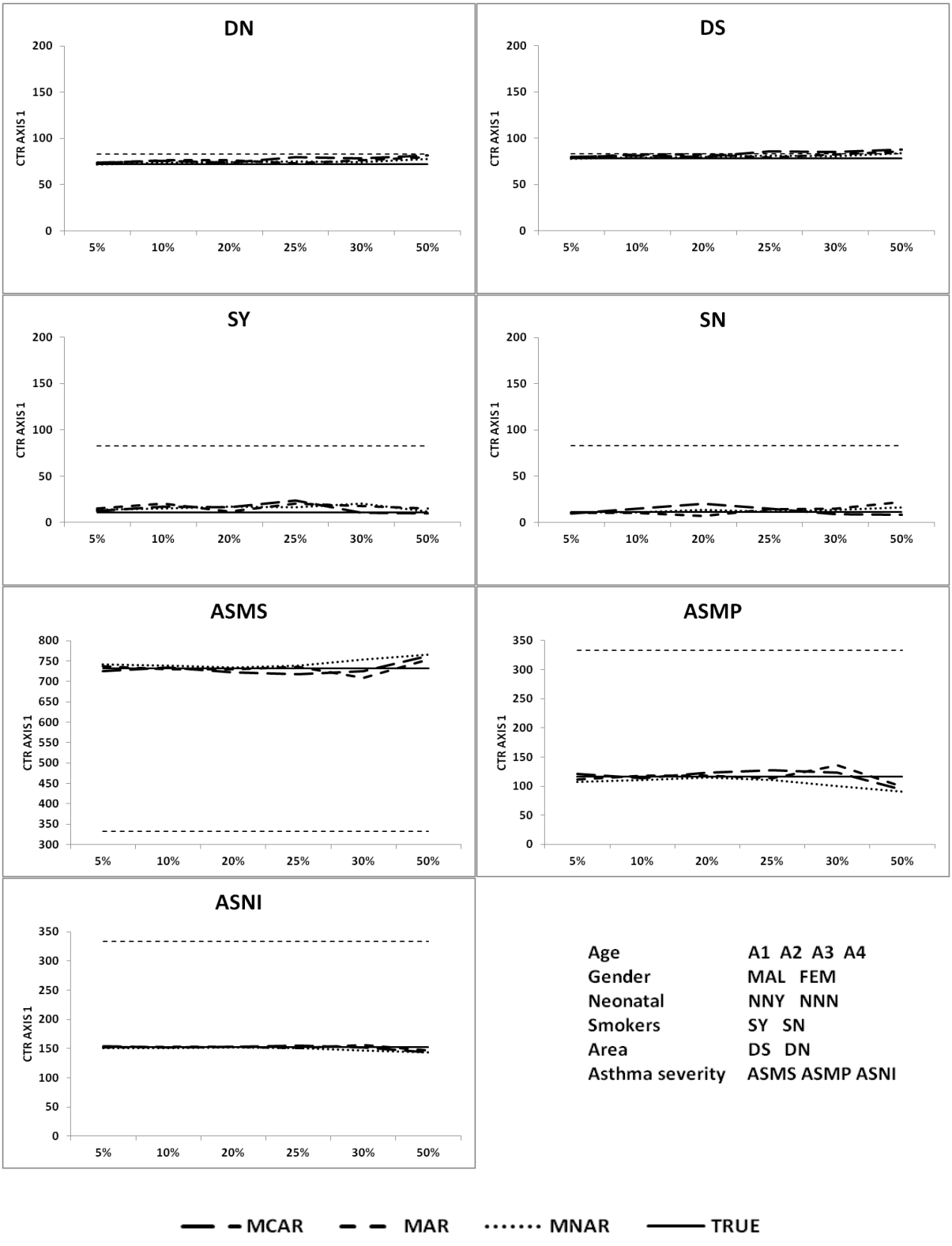



**Figure 3(continued): CTR values for all variables on axis 1 (Dotted line indicates the threshold value)**

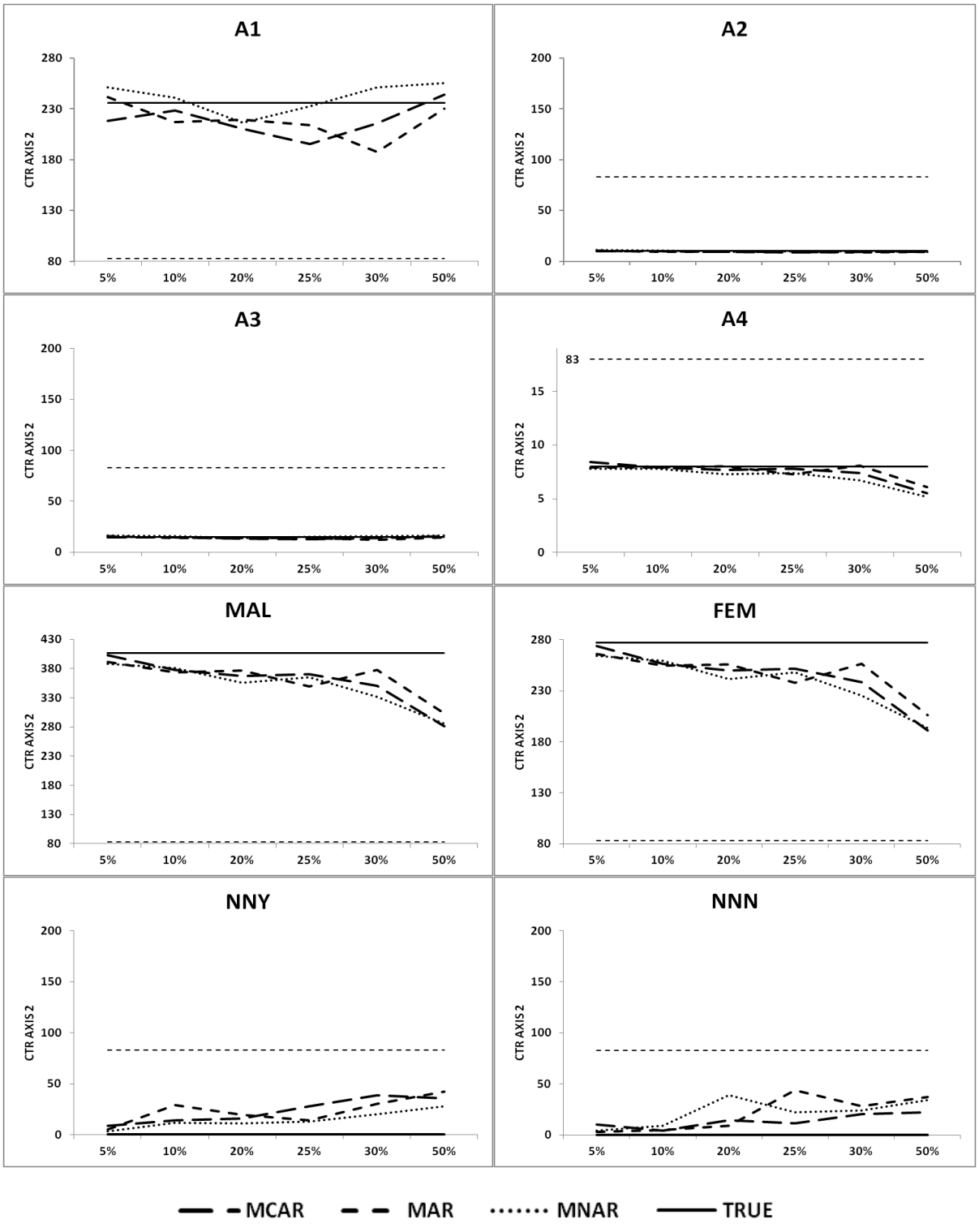



**Figure 4: CTR values for all variables on axis 2 (Dotted line indicates the threshold value)**

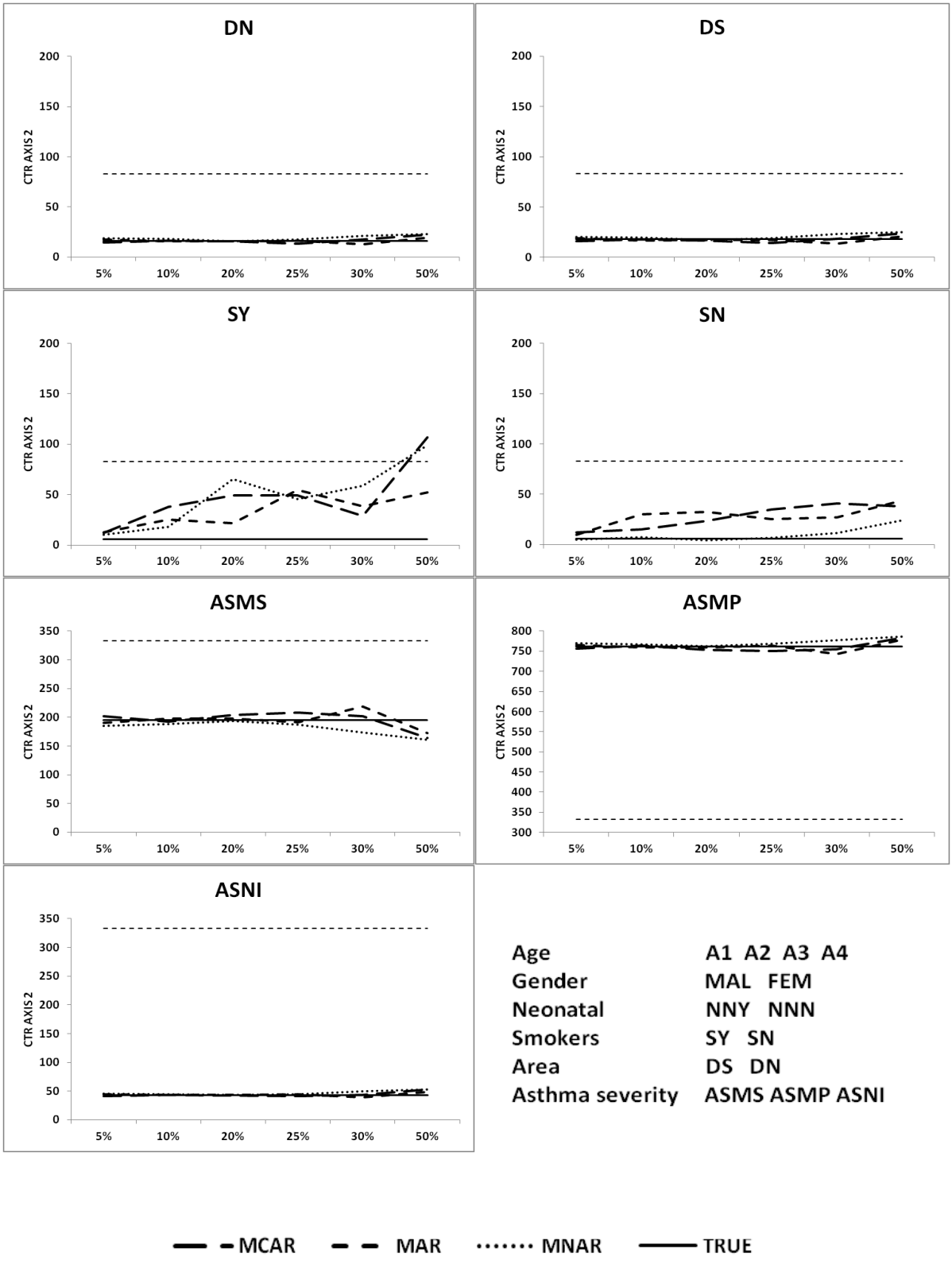



**Figure 4(continued): CTR values for all variables on axis 2 (Dotted line indicates the threshold value)**

CTR values, which have been scaled to sum to 1000 for each axis, indicate the amount that each variable contributes to the inertia of the axis. As a rule of thumb, if the CTR value of a point exceeds the average contributions of all the points (rows or columns) for a particular axis, then that point can be considered important to the orientation of the axis and is used in the interpretation of the results. In this study, an approximate CTR threshold value for the asthma severity categories is 333 and for the other variable categories it is 83.

**AXIS 1** - contrasts lowest asthma severity category (ASNI) with higher asthma severity categories (ASMP and ASMS)

Across all scenarios of MM and M%, A1, MAL, NNY and ASMS remain above the threshold value of importance while A2, A3, A4, NNN, SY, SN, DN, ASNI and ASMP remain below the threshold value of importance to this axis. Both FEM and DS are marginal with FEM positioned just above the threshold value and DS hovering around that value.

No significant differences between MM's were found for any variable category except NNN where, at 30% and 50%, CTR values for MAR are significantly higher than MNAR and MCAR values respectively. Furthermore, the only significant differences across M% were found for MAL, FEM and A4 where the CTR values increase significantly at 50% for all MM's.

**AXIS 2** – contrasts the two highest asthma severity categories: ASMS vs ASMP

All variable categories remain distinctly positioned relative to the threshold values except for SY which crosses the threshold at 50% for the MCAR and MNAR mechanisms. The only significant difference between MM's across all scenarios is at 25% on SN where the CTR value for MCAR is significantly higher than the value for MNAR. The only significant differences across M%'s were found for MAL, FEM and A4 where CTR values decrease significantly at 50% for all MM's.

*Model inertia values*

The total inertia of the measured data does not change significantly across either MM or M% (Figure 5). However, the total inertia of the measured data, taken as a percentage of the total inertia of the full data set – measured and missing – is significantly higher for MCAR at 50% than at 5%. There is no significant difference in this measure across MM (Figure 5).

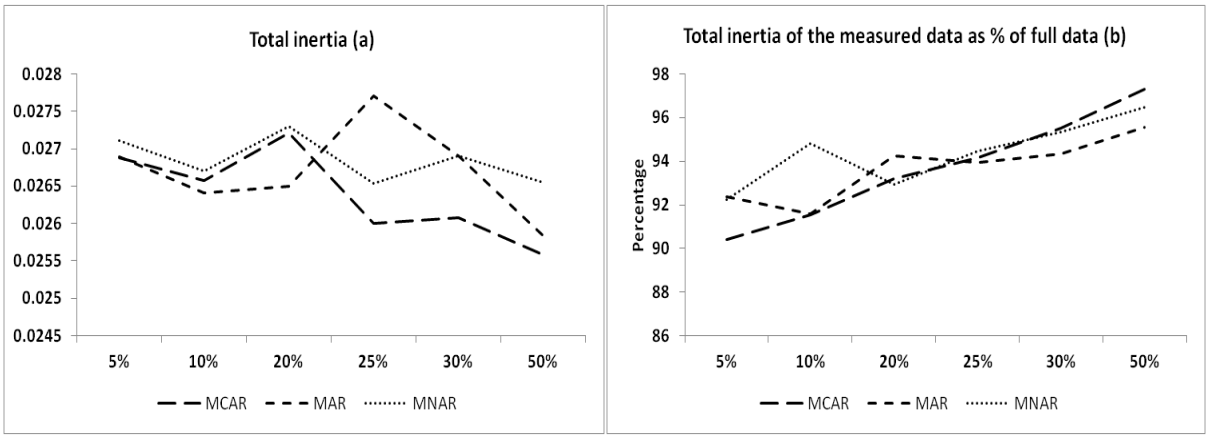



**Figure 5: Measures of Inertia**

*Graphical displays*

A degree of 'movement' is evident for some of the variable categories in the display of the subspace define by axis 1 and axis 2 (Figure 6). This dispersion is more evident in the variables that have undergone missingness. For the variables that are fully measured, dispersion appears to be greater in those variable categories that have more variability. The variables that are further from their true positions are those with higher percentage missingness, with no specific correlation to mechanism.

## Discussion

The aim of this simulation study was to explore the effect of both the missingness mechanism and the amount of missingness present in data on the use of sCA as applied to categorical data that suffers from missingness.

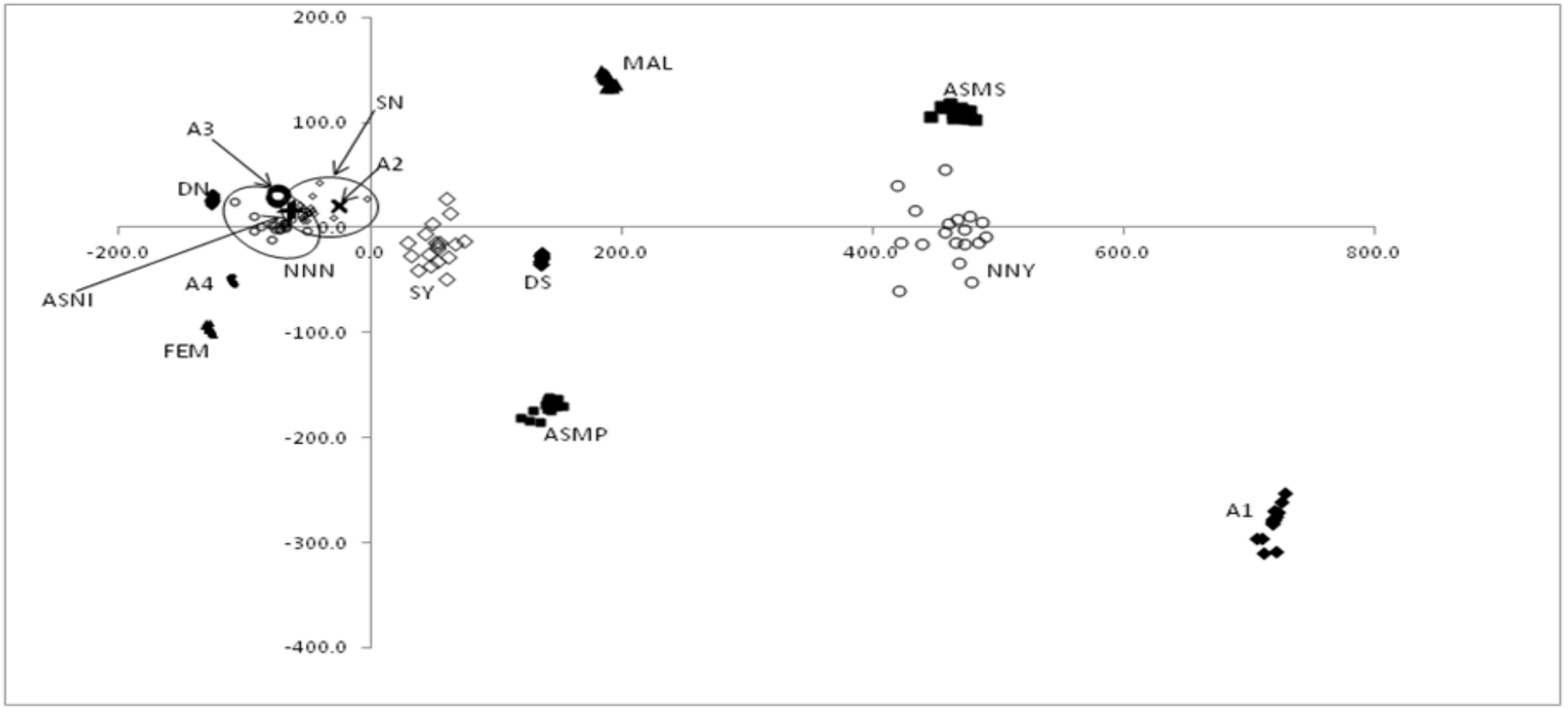



**Figure 6: Graphical representation of all variables for all scenarios**.

| Key to variables: | **Age** | A1/A2/A3/A4 | **Area** | DS/DN |
|---|---|---|---|---|
| | **Gender** | MAL/FEM | **Smokers** | SY/SN |
| | **Neonatal** | NNY/NNN | **Asthma severity** | ASNI/ASMP/ASMS |

Three accepted MM's (MCAR, MAR and MNAR) were simulated, each across six degrees of missingness (5%, 10%, 20%, 25%, 30% and 50%).To allow for sampling variability, data for each of these 18 scenarios was generated 10 times and results were averaged.

It was found that the MM did not substantiallty affect the results. Furthermore, for missingness of up to 25% per variable, results were not notably affected.

The effect of the MM on all outcomes is negligible. In terms of the COR outcome, the only effect of the MM is on the variable category SN. In this case, MNAR values are significantly different from MCAR and MAR values but are in fact closer to the 'true' values.

Only one incident of significant difference across the MM's in CTR values for axis 1 was found. For the variable category NNN, CTR values for MAR are greater than those for MNAR and MCAR at 30% and 50% respectively. Likewise, the single significant difference across MM in CTR values for axis 2 was found for the variable category SN at 25% where the CTR value for MNAR is significantly smaller that the value for MCAR, but in fact also closer to the 'true' value. Since the position of these measures relative to the threshold values remains the same, the differences do not affect the interpretation and these variable categories remain trivial to the orientation of axes 1 and 2.

These aforementioned differences across mechanisms all occur in the variable categories NNN and SN, which have been subjected to missingness but which are low in importance in terms of the orientation of the axes. The other variable categories that underwent missingness, SY and NNY, do not experience any significant differences across MM for any of the measures.

Greater deviations from the true COR values as well as variations in COR values across missingness mechanisms are apparent for the variable categories that underwent missingness. For each of these four variable categories, COR values are consistently lower than the 'true' values. In the case of SY and SN, COR values decrease from 910 to below 500 for some scenarios,

thus indicating that axis 1 is no longer the most important axis to these points for all scenarios. This suggests that the importance of an axis to the inertia of a point decreases when missingness is introduced and can result in a change in the axis that contributes most to the inertia of a point. So the point can, at times, 'hop axes'. However, it is clear from the graphical display of the points that even when there is a drastic reduction in COR value, resulting in a possible 'hopping' to another axis, the final placement of the point in the subspace is not compromised .

An effect of M% is found on COR values for the variable categories NNN, SY and SN. In the case of NNN, values of COR generally deviate more from the 'true' values as the percentage missing increased and there is a significant drop in COR value for MCAR at the 20% missing stage. A similar scenario exists for SN where there is a significant decrease in the COR value for MCAR as the amount of missingness increases. In the same way, the COR value for MNAR on SY is significantly lower at 50% missingness than at 5% missingness. The effect of M% on the asthma severity categories, ASNI and ASMP, indicates that there is a significant decrease in COR values at 50% missingness.

For those variables that did not undergo missingness, COR values across the three mechanisms do not differ appreciably from the true values for up to at least 25% missingness. Some differences, while not significant, are apparent for 30% and 50% missingness. Variables with missingness show erratic deviations from true values across all scenarios. These are especially pronounced in the variable categories that do not contribute appreciably to the general analysis.

The only significant effects of M% on CTR values for both axes are the significant change in values at 50% missingness for A4, MAL and FEM.

Apart from deviations from true CTR values being experienced by some incomplete variables, some deviation is also apparent in completely measured variables whose CTR values lie above the importance threshold. In no instance does this affect the overall outcome and interpretation.

While changes in the CTR values, in general, do not make a difference to the importance of points to the inertia of axes 1 and 2, there is one exception. CTR values for SY on axis 2 have increased from below to above the threshold level at 50% missingness for two of the MM's – MCAR and MNAR. This indicates that as missingness increases, it is possible for a variable to *become* 'significantly' important to the inertia of an axis. It must be remembered that, under the MNAR mechanism, data was deleted from the smoking variable at a ratio of 90:10 for SY:SN. Thus, compared to SN, a large proportion of data from SY would have been missing at 50% missingness. This would account for the greater effect of M% on SY than on SN under the MNAR mechanism and could indicate that under extreme missingness, analyses can lose some stability.

Total inertia for the measured data, a measure of the variability in the measured data, is not significantly affected by either the MM or the degree of missingness. However, when total inertia of the measured data is taken as a percentage of the total inertia of the full data set including the missing data, the percentage missing has an effect. For MCAR, this measure at 50% missingness is significantly higher than at 5% missingness. Visually, there is an upward trend across all mechanisms as missingness increases. This may imply that the variability in the missing data is significantly lower for the MCAR mechanism at 50% missingness than at 5% missingness.

Examining the plot of all variables across the 18 scenarios confirms that there is some deviation from the true position for some variables. This is most evident in the variables that suffer from missingness but is also present in the completely measured variables that show stronger associations with asthma severity (MAL and A1). In general, while points that are further from their 'true' position have a higher percentage missing (not shown on the plot), there is no evidence that the MM is a factor in this displacement. In all cases, the dispersion is well contained and the relative positioning of variable categories with each other and with the axes remains unchanged.

Results from this study suggest that there is no evidence that the MM has a negative effect on results after applying sCA to data that suffers from missingness. In addition, the analysis has shown that only when missingness exceeds the 30% level, are some results affected. However, while deviations in the outcomes studied are present, they do not affect the overall interpretation of the analysis.

*Limitations*

While we are confident that the results emanating from this study are reliable, there are some limitations. These results are specific to the variables included and the mechanisms imposed on this data. The variables were selected according to their relationships with asthma severity such that all strengths of relationship are represented. Three of the four variable categories that underwent missingness do not have strong relationships with asthma severity. We chose to impose missingness on only two of the six variables. Furthermore, while the deletions on these variables were based on plausible judgements, they are subjective, and may have influenced the findings. Further studies need to be carried out to explore the effect of different ratios of missingness and a different or increased choice of variables.

We performed 10 simulations on our data. The number of simulations to perform is dependent on the required accuracy with an increase in the number of simulations resulting in more accuracy (Burton, Altman, Royston, & Holder, 2006; Ritter, Schoelles, Quigley, & Klein, 2011). In contrast to confirmatory techniques, in which relationships are hypothesised and proved, CA (and its variants) is an exploratory approach in which relationships in the data are revealed and visualized for purposes of interpretation. Relative positions of category points indicate levels of similarity or association between categories. No measures of statistical significance are applied (Greenacre, 1992). Thus accuracy is not of prime importance. There is little evidence to suggest that additional simulations would have produced meaningfully different results.

## Conclusions

Under the conditions imposed in this study, we found that there is no evidence to suggest that the missingness mechanism has an effect on results when sCA is applied to data that suffers from missingness. It was found that, in some cases, values of the outcomes studied deviate from the true values when the amount of missingness exceeds 30% per variable. These deviations do not, however, affect the overall interpretation of the results. We believe that sCA would have a similar impact on other data sets that comprise categorical variables that suffer from missingness.

## References


Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine, 25*(24), 4279-4292.

Donders, A. R. T., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: a gentle introduction to imputation of missing values. *Journal of clinical epidemiology, 59*(10), 1087-1091.

Gelman, A., & Hill, J. (2007). Missing-data imputation. *Behavior research methods, 43*(2), 310-330.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology, 60*, 549-576.

Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. (1997). Analysis with missing data in prevention research. *The science of prevention: Methodological advances from alcohol and substance abuse research, 1*, 325-366.

Greenacre, M. (1992). Correspondence analysis in medical research. *Statistical methods in medical research, 1*(1), 97-117.

Greenacre, M., & Pardo, R. (2006). Multiple correspondence analysis of a subset of response categories. In M. Greenacre & J. Blasius (Eds.), *Multiple Correspondence Analysis and Related Methods* (pp. 197 - 218). Boca Raton: Chapman & Hall/CRC

Greenland, S., & Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American journal of epidemiology, 142*(12), 1255-1264.

Hendry, G., North, D., Zewotir, T., & Naidoo, R. (2014). The application of subset correspondence analysis to address the problem of missing data in a study on asthma severity in childhood. *Statistics in medicine, 33*(22), 3882-3893.

Little, R., & Rubin, D. (1987). Statistical Analysis with Missing Data. *John A. Wiley & Sons, Inc., New York*.

Marshall, A., Altman, D. G., Royston, P., & Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC medical research methodology, 10*(1), 7.

Naidoo, R. N., Robins, T. G., Batterman, S., Mentz, G., & Jack, C. (2013). Ambient pollution and respiratory outcomes among schoolchildren in Durban, South Africa. *South African Journal of Child Health, 7*(4), 127-134.

Peyre, H., Leplège, A., & Coste, J. (2011). Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of Life Research, 20*(2), 287-300.

Ritter, F. E., Schoelles, M. J., Quigley, K. S., & Klein, L. C. (2011). Determining the number of simulation runs: Treating simulations as theories by not sampling their behavior *Human-in-the-Loop Simulations* (pp. 97-116): Springer

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81): John Wiley & Sons.

Scheffer, J. (2002). Dealing with missing data.

Shrive, F. M., Stuart, H., Quan, H., & Ghali, W. A. (2006). Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. *BMC Medical Research Methodology, 6*(1), 57.

Vach, W., & Blettner, M. (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *Am J Epidemiol, 134*(8), 895-907.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research, 16*(3), 219-242.

Wayman, J. C. (2003). *Multiple imputation for missing data: What is it and how can I use it.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

# A review on the application of multiple imputation and subset correspondence analysis in the South Durban health study

Gillian M. Hendry*[1]

1 Corresponding author

Email: hendryfam@telkomsa.net

Temesgen Zewotir*

Email: Zewotir@ukzn.ac.za

Rajen N. Naidoo**

Email: naidoon@ukzn.ac.za

Delia North*

Email: northd@ukzn.ac.za

* School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville Campus, University Road, Westville, Durban, South Africa

** Discipline of Occupational and Environmental Health, School of Nursing and Public Health, University of KwaZulu-Natal, Durban, South Africa

## Abstract

### Background

Missing data is commonly encountered in most studies where data is collected by means of surveys. Furthermore, it is likely that much of the data will be categorical. Among the methods, most frequently found in literature, that are used to handle the missing data are complete case analysis and mean substitution. These, and other ad hoc methods of dealing with missing data, however, can lead to biased estimates and are not, in general, recommended. Recently, there has been a move towards applying multiple imputation to take care of the non-response. This method has been shown to produce reliable parameter estimates under most conditions. Another, less well-known method that takes a very different approach but that effectively deals with the missing data, is subset correspondence analysis.

### Methods

In this paper, multiple imputation and subset correspondence analysis are applied to a set of child asthma data that is mainly categorical and suffers from non-response. Differences in the methods and in the outcomes they produce are studied. In addition, the inclusion of interactions in a subset correspondence analysis is illustrated.

### Results

Despite the vast differences in the two approaches, they yielded similar results in the identification of genetic, environmental and socio-economic factors that affect childhood asthma. A number of exposure related variables were found to be

associated with the greater severity of asthma. It was also found that a finer distinction between the asthma severity levels and their associations with factors was possible with a subset correspondence analysis, compared to the multiple imputation approach.

**Conclusions**

Both multiple imputation and subset correspondence analysis were able to identify several factors associated with childhood asthma while at the same time successfully managing the missing data. This offers the researcher a choice to select the method that best suits his/her study.

## Background

The collection of data by means of surveys generally elicits some non-response resulting in missing data. Depending on the reason for the non-response, the missingness can be classified according to the popular definitions suggested by Little and Rubin [1]. Missing values that do not depend on either observed or unobserved data are termed “missing completely at random” (MCAR); if the missing values are independent of unobserved data but may depend on observed data, they are referred to as “missing at random” (MAR); and missing values that depend on both observed and unobserved data are termed “missing not at random” (MNAR).

It has been common practice to deal with the missing data by applying any of a number of ad hoc methods. These include, amongst others, mean substitution, hot

deck imputation, the indicator method and pairwise deletion. The most commonly applied method, however, is case-wise deletion, also called complete case analysis [2] , in which all cases with missing data items are dropped from the analysis. This can result in a significantly reduced sample size or bias and can negatively affect results. In most instances, unless the data is MCAR, the application of these aforementioned ad hoc methods, whether they impute missing values or drop cases, result in biased estimates and are therefore not recommended [1, 3]. In reality, the missingness mechanism present in the data is rarely solely MCAR but rather a combination of mechanisms and so another means of dealing with the missing data is required.

More recently much work has been done on the development of multiple imputation methods to deal with missing data. The concept of multiple imputation was first introduced in the late 1970's by Rubin [4]. However, due to the computationally intense nature of the process and the absence of sufficiently powerful computers, the application of multiple imputation did not take off until the 1990's with the advent of computers with enhanced computational capabilities. Several algorithms have been developed and efficient software is now more freely available.

Two algorithms that are widely available and frequently used for imputation when missing data occurs in a general pattern (nonmonotonic missingness) are: multiple imputation based on the multivariate normal distribution (MVNI), available in a standalone package – NORM – developed by Schafer [5]; and an algorithm known as "fully conditional specification" (FCS) or "chained equations" – implemented by, amongst others, van Buuren et al [6] - available in a number of commercially available

statistical packages. FCS is more flexible than MVNI in that it does not depend on the assumption of multivariate normality and is applicable to a mix of variable type.

Because a large proportion of the non-response in survey data is often found in categorical variables, this paper addresses, in particular, the problem of exploring the relationships between categorical variables that suffer from missingness. We used the FCS approach to multiple imputation to deal with the missing data and then completed the analysis by applying ordinal regression to the imputed data sets.

In contrast to this aforementioned approach which includes traditionally accepted methodology classically favoured by epidemiologists, subset correspondence analysis (s-CA) is also effective in exploring relationships between categorical variables and at the same time taking care of the missing data. s-CA is a variant of correspondence analysis (CA) and was developed by Greenacre and Pardo [7]. It involves the application of CA to a subset of the data. In its application to incomplete data, the non-response for each variable is categorized separately and CA is applied to the subset of observed categories.

These two methods adopt different philosophies in their approach to analysis and whereas the one is governed by distributional requirements and missingness mechanisms, the other is not. While the application of both methods to the analysis of missing data has been illustrated [8, 9] , no comparison has yet been made. Furthermore, the inclusion of interactions in the application of s-CA with missing data is not evident in the literature.

In this paper, we compare the use of these two somewhat different methods on a set of epidemiological data, with a large number of categorical variables in which missingness was present, from a study of asthma severity in children in Durban, South Africa. We also illustrate the inclusion of interactions in these analyses.

## Methods

The motivating problem for this investigation was the analysis and reporting of the respiratory health of children in the South Durban region of KwaZulu-Natal, South Africa. The data (Table 1) includes information from 382 children on 17 environmental, socio-economic, genetic and behavioural variables as well as a three-tiered asthma severity measure. All but one of the variables – age – are categorical. Of the 382 subjects, 27 (7.1%) were classified as having moderate to severe asthma; 47 (12.3%) suffered from mild persistent asthma; and the remaining 308 (80.6%) either showed symptoms for possible asthma or did not exhibit definite asthma symptoms. This data set is potentially rich in its ability to reveal relationships between the outcome variable asthma severity, an ordinal measure, and the environmental, genetic, socio-economic and behavioural variables. However, data amounting to 5.1 % of the total is missing from the data set. This is spread across 43.5% of the 382 records, thus leaving only 216 complete records. A standard approach when seeing these data might be to run an ordinal logistic regression of asthma with the logit link function. However, the standard logistic regression estimation methods require complete data. Consequently, cases with incomplete data are ignored, leading to bias when data are MNAR or MAR, and a loss of power when data are MCAR.

Table1: Categories, code names and frequencies for all variables

| Variables | Categories (code names) – count (N = 382) | | | | | | | | Non-response – count (%) |
|---|---|---|---|---|---|---|---|---|---|
| *Gender* | male (male) | 163 | female (fem) | 219 | | | | | 0 |
| *Exercise* | <twice weekly (E1) | 113 | 2-4 times/wk (E2) | 135 | >4 times/wk (E3) | 110 | | | E* - 24(6) |
| *TV watching* | <1 hr a day (T1) | 86 | 1 - 3 hours/day (T2) | 193 | > 3 hours/day (T3) | 78 | | | T* - 25(7) |
| *Smokers in the home* | yes (SY) | 187 | no (SN) | 194 | | | | | Sm* - 1(<1) |
| *Breakfast habits* | daily (BD) | 236 | not daily (BN) | 121 | | | | | B* - 25(7) |
| *Pets at home* | yes (PY) | 114 | no (PN) | 264 | | | | | P* - 4(1) |
| *Food availability* | enough food (Fe) | 265 | not enough (Fn) | 85 | | | | | F* - 32(8) |
| *Work and wear* | yes (WWY) | 36 | no (WWN) | 332 | | | | | WW* -14(4) |
| *Smoke while pregnant* | yes (SPY) | 35 | no (SPN) | 328 | | | | | SP*-19(5) |
| *Neonatal care* | yes (NY) | 50 | no (NN) | 318 | | | | | N*-14(4) |
| *Fear in neighbourhood* | yes (FrY) | 165 | no (FrN) | 192 | | | | | Fr* - 25(7) |
| *Violence experienced* | yes (VY) | 185 | no (VN) | 169 | | | | | V* - 28(7) |
| *Smokers in vehicles* | yes (SVY) | 94 | no (SVN) | 259 | | | | | SV*-29(8) |
| *Num of people in home* | 1 - 4 people (Np1) | 124 | 5 - 7 people (Np2) | 153 | >7 people (Np3) | 70 | | | Np* - 35(9) |
| *Age** | 8-9 yrs(A1) | 25 | 10 years(A2) | 196 | 11 years(A3) | 135 | 12+years(A4) | 26 | 0 |
| *Income* | Up to R1000(I1) | 79 | R1001 – R4500(I2) | 102 | R4501– R10000(I3) | 88 | R100001+ (I4) | 39 | I* - 74(19) |
| *Area* | South Durban(SD) | 197 | North Durban(ND) | 195 | | | | | 0 |
| *Asthma severity* | Moderate/severe(ASMS) | 27 | Mild persistent(ASMP) | 47 | Probable/no.(ASPN) | 308 | | | 0 |

* also available as a scale variable

This table is modified from Table 1 in [9]

**Existing approaches for handling missing data**

Two methods that have previously been successfully applied to this data set are multiple imputation followed by ordinal regression and s-CA [8, 9].

The multiple imputation process involves three basic steps: “filling in” the missing values with reasonable predictions multiple times, creating multiple complete data sets; separately analysing each of the imputed data sets; and combining the results according to Rubin’s rules [10].

With the application of the FCS approach to multiple imputation, a series of regression models are run such that each variable with missing data is regressed on the other variables according to its distribution. In particular, categorical variables are modelled using logistic regression. This is an iterative process that is repeated until parameters from the regression model have stabilized at which time one complete data set is produced. The entire process is repeated until the required number of imputed data sets is generated. The analysis of the imputed data sets followed by the combining of the results identifies the strength of the relationships between the independent variables and the dependent variable. Details of this iterative method can be found in Azur et al [11].

In contrast, CA is a graphical technique used in the analysis of categorical data. While the more classical regression-based methods for studying inter-variable relationships hypothesise a model and fit the data to the model, CA does not hypothesise a model but rather decomposes the data in order to study their structure [12]. Rows and columns of a rectangular data matrix, which represent points in multidimensional

space, are optimally displayed in a lower dimensional subspace thus enabling the interpretation of relationships between the variables.

CA is usually applied to a full data set. However, with the development of s-CA, it is possible to effectively manage the missing data without losing any of the measured data. A more detailed description of this method as applied to incomplete data can be found in Hendry et al [9] .

**Methodologies adopted for this comparative study**

In order to understand the comparative strategies it is essential to understand how differently the two processes (Multiple imputation and s-CA) operate when identifying the factors associated with child asthmatic levels in the presence missingness.

In the multiple imputation approach, based on the amount of missingness present, 20 data sets were imputed [13]. To ensure stability of the parameters, ten iterations of the imputation process were completed between each retained complete data set [14].

Each of the 20 imputed data sets was analysed using ordinal regression with the logit link function. The results were then combined following Rubin's rules [15]. Overall parameter estimates were calculated as the average of the parameter estimates obtained from the analysis of each data set; and the variances of the overall parameter estimates were calculated as a function of both the variance within each data set and the variance across the data sets.

Both the multiple imputation and the analysis of the imputed data sets were carried out using the Statistical Package for Social Sciences (SPSS version 17).

Before imputing missing values, it was necessary to carry out tests in order to identify the missingness mechanism present in the data. For each incomplete variable, an indicator variable was created and chi-square analyses were performed to test whether either the incomplete variable or its missingness was related to observed values of other variables. This enabled the identification of variables necessary to include in the imputation model in order to make the MAR assumption as plausible as possible [13] .

The identification of interactions, in the presence of missing data, presents a challenge [16, 17] . In a previous study using this data set, this problem was addressed and 10 interactions were identified as being significant [8]. For the purposes of this study, the two strongest interactions – 'gender * smoke exposure in vehicles' and 'fear * breakfast habits' – were included in the analysis. Interaction product terms were coded into separate categories and treated as additional variables in the imputation model. For example: the interaction gender (male/female) * smoke exposure in vehicles (yes/no) was broken down and coded as male/yes = 1; male/no = 2; female/yes = 3; female/no = 4. The interaction categories, along with the remaining 17 variables – one scale and 16 categorical – were treated as predictor variables.

On the hand, the s-CA approach deals with the objective of identifying the association of environmental, genetic, behavioural and socio-economic variables with asthma severity, by re-organising the data in the form of a contingency table. The columns

represent the three asthma severity categories and the rows represent the categories of the 17 variables and two interactions, with the interactions broken down and coded as for the imputation model.

For the purpose of this analysis, the variable 'age' was classified into 4 categories. To manage the missing data, a 'missing' category was introduced for each variable with missing data. The subset to be analysed was formed by excluding these missing categories.

Variables involved in the interactions – 'gender', ' smoke exposure in vehicles', 'fear' and 'breakfast habits' - were not included as individual active variables in the analysis but were treated as supplementary variables [12]. By so doing, they do not participate in the orientation of the axes but their individual positions as "main effects" relative to the associated interactions [18] can still be studied.

A macro program was written to perform the s-CA.

As seen above the two approaches have no common parameter estimates or model structure. Thus the conventional comparison of methods in terms of mean square errors or goodness of fit is not directly applicable. Accordingly a systematic holistic review of the two approaches is adopted.

## Results and discussion

The aim of this study was to illustrate and compare two methods to analyse categorical data that suffers from missingness. We found that, while multiple imputation, in combination with ordinal regression, and CA applied to a subset of data

are vastly different methodologically, the results that they produce in the analysis of inter-variable relationships are very similar.

The application of these two methods enabled us to identify relationships between asthma severity and several environmental, genetic, socio-economic and behavioural variables and, at the same time, retain all records. Furthermore the associations between these variables and asthma (Table 2 and Table 3) were consistent across methods and generally confirmed established theories regarding factors that exacerbate asthma. There was agreement that confirmed asthma is associated with children who: are younger [19]; have had some special neonatal care [20]; are exposed to smoke in the home [21, 22], in vehicles [23], *in utero* [24] and in the form of air pollution [25, 26]; lived in a home with up to 4 people [27]; come from a R4501 – R10000 income household; do not always have enough food; are exposed to low concentrations of compounds and pollutants [28, 29]; never had a pet and do not experience fear in the neighbourhood. Both analyses also indicated an association between worse asthma and both lack of violence in the neighbourhood and watching up to one hour of TV a day. These associations are contrary to what other studies have found and, while the data was explored for reasons for these anomalies, none were found. We concluded that there must be some underlying factor specific to this sample.

Table 2: Estimated Coefficients (EST) and Standard Errors (SE)

| Predictor | Reference Category | Category | FCS(N = 382) | |
|---|---|---|---|---|
| | | | EST | SE |
| Gender | Female | Male | 0.039 | 0.362 |
| Neonatal care | No | Yes | 0.847* | 0.394 |
| Fear | No | Yes | -1.042* | 0.406 |
| Smoked while pregnant | No | Yes | 0.379 | 0.488 |
| Smokers in home | No | Yes | 0.701* | 0.309 |
| Smoke in vehicles | No | Yes | -0.706 | 0.512 |
| Exercise | >4 times a week | Up to once a week | 0.044 | 0.384 |
| | | 2 – 4 times a week | 0.011 | 0.384 |
| TV watching | >3 hours a day | Up to 1 hour a day | 0.786 | 0.465 |
| | | 1 – 3 hours a day | 0.046 | 0.43 |
| Number people in home | 8+ | 1 - 4 | 0.981* | 0.481 |
| | | 5 - 7 | 0.381 | 0.494 |
| Income | R100001+ | up to R1000 | -0.133 | 0.611 |
| | | R1001 – R4500 | 0.017 | 0.555 |
| | | R4501 – R10000 | 0.697 | 0.523 |
| Food availability | Enough | Not always enough | 0.756 | 0.406 |
| Work'nWear | No | Yes | 0.402 | 0.456 |
| Pets ever | No | Yes | -1.072* | 0.398 |
| Area | North Durban | South Durban | 0.595 | 0.306 |
| Breakfast habits | Daily | Not daily | -1.011* | 0.494 |
| Violence | No | Yes | -0.709* | 0.34 |
| Age | | | -0.247 | 0.16 |
| Fear * Breakfast | No/Daily | Yes/Not daily | 2.338* | 0.725 |
| Gender * SmokeVehicle | Female/No | Male/Yes | 1.811* | 0.699 |

ND – North Durban; SD – South Durban; preg – pregnant;

FCS -Multiple imputed FCS

*Significant at the 0.05 level

Table 3: Decomposition of inertia for the 2 principal axes

| Name | Mass | INR | k= 1 | COR | CTR | k= 2 | COR | CTR |
|---|---|---|---|---|---|---|---|---|
| A1 | 4 | 3 | -717 | 955 | 146 | -156 | 45 | 42 |
| A2 | 34 | 0 | 34 | 754 | 3 | 19 | 246 | 5 |
| A3 | 24 | 0 | 73 | 874 | 8 | 28 | 126 | 7 |
| A4 | 5 | 0 | 58 | 151 | 1 | -138 | 849 | 34 |
| NNY | 9 | 3 | -476 | 985 | 129 | 59 | 15 | 12 |
| NNN | 55 | 0 | 66 | 1000 | 16 | -1 | 0 | 0 |
| SPY | 6 | 3 | 13 | 27 | 0 | -75 | 973 | 14 |
| SPN | 57 | 57 | -7 | 454 | 0 | 7 | 546 | 1 |
| SY | 33 | 23 | -51 | 969 | 6 | -9 | 31 | 1 |
| SN | 34 | 34 | 47 | 972 | 5 | 8 | 28 | 1 |
| E1 | 20 | 11 | -2 | 480 | 0 | -2 | 520 | 0 |
| E2 | 24 | 6 | 60 | 667 | 6 | -42 | 333 | 17 |
| E3 | 19 | 24 | -14 | 62 | 0 | 56 | 938 | 24 |
| T1 | 15 | 2 | -186 | 454 | 34 | 204 | 546 | 248 |
| T2 | 34 | 4 | 45 | 326 | 4 | -65 | 674 | 56 |
| T3 | 14 | 3 | 147 | 997 | 19 | -8 | 3 | 0 |
| N1 | 22 | 18 | -170 | 994 | 41 | -13 | 6 | 1 |
| N2 | 27 | 7 | 61 | 983 | 7 | 8 | 17 | 1 |
| N3 | 12 | 48 | 160 | 934 | 21 | 43 | 66 | 9 |
| I1 | 14 | 0 | 41 | 952 | 2 | 9 | 48 | 0 |
| I2 | 18 | 13 | 33 | 501 | 1 | -33 | 499 | 8 |
| I3 | 15 | 4 | -191 | 992 | 37 | 18 | 8 | 2 |
| I4 | 7 | 73 | 102 | 977 | 5 | 16 | 23 | 1 |
| Fn | 46 | 31 | 59 | 938 | 11 | 15 | 62 | 4 |
| Fe | 15 | 6 | -31 | 435 | 1 | 35 | 565 | 7 |
| WWY | 6 | 11 | -406 | 915 | 68 | -124 | 85 | 38 |
| WWN | 58 | 35 | 35 | 736 | 5 | 21 | 264 | 10 |
| PY | 20 | 2 | 199 | 951 | 51 | 45 | 49 | 16 |
| PN | 46 | 42 | -80 | 998 | 19 | -3 | 2 | 0 |
| DS | 33 | 15 | -125 | 997 | 33 | -6 | 3 | 1 |
| DN | 34 | 6 | 120 | 997 | 32 | 6 | 3 | 1 |
| VY | 32 | 12 | 111 | 991 | 26 | 10 | 9 | 1 |
| VN | 29 | 25 | -95 | 978 | 17 | -14 | 22 | 2 |

Mass (Mass) and inertia (INR) of each variable; the quality (QLT) of the variable's representation in the subspace of the first 2 axes; Co-ordinates (k = ...); contributions of axes to the inertia of the variables (COR); and contributions of variables to the inertia of the axes (CTR)*.

Table 3: Decomposition of inertia for the 2 principal axes (continued)

| Name | Mass | INR | k= 1 | COR | CTR | k= 2 | COR | CTR |
|---|---|---|---|---|---|---|---|---|
| FYBd | 20 | 12 | 179 | 987 | 41 | 20 | 13 | 3 |
| FNBd | 21 | 78 | -203 | 931 | 58 | -55 | 69 | 26 |
| FYBn | 9 | 6 | -141 | 654 | 12 | 103 | 346 | 38 |
| rNBn | 12 | 9 | 228 | 957 | 40 | 48 | 43 | 11 |
| mSVY | 7 | 276 | -363 | 998 | 60 | 16 | 2 | 1 |
| mSVN | 19 | 22 | -76 | 158 | 7 | 176 | 842 | 228 |
| fSVY | 9 | 74 | 137 | 947 | 12 | -32 | 53 | 4 |
| fSVN | 27 | 3 | 104 | 480 | 19 | -109 | 520 | 124 |
| | | | | | | | | |
| ASMS | 71 | 89 | -371 | 929 | 635 | 103 | 71 | 295 |
| ASMP | 123 | 130 | -157 | 638 | 198 | -118 | 362 | 679 |
| ASNI | 806 | 781 | 56 | 975 | 168 | 9 | 25 | 26 |
| **SUPPLEMENTARY** | | | | | | | | |
| MAL | | | -150 | 516 | | 146 | 484 | |
| FEM | | | 107 | 494 | | -108 | 506 | |
| FrY | | | 66 | 808 | | 32 | 192 | |
| FrN | | | -50 | 896 | | 17 | 104 | |
| SVY | | | -79 | 980 | | -11 | 20 | |
| SVN | | | 27 | 894 | | 9 | 106 | |
| Bnd | | | 68 | 462 | | 73 | 538 | |
| Bd | | | -22 | 596 | | -18 | 404 | |

Mass (Mass) and inertia (INR) of each variable; the quality (QLT) of the variable's representation in the subspace of the first 2 axes; Co-ordinates (k = ...); contributions of axes to the inertia of the variables (COR); and contributions of variables to the inertia of the axes (CTR)*.

The interpretation of the interactions was also consistent across methods. With regard to the 'gender * smoke exposure in vehicle' interaction, it was found that male children who are exposed to smoke in a vehicle suffer from significantly worse asthma than girls not exposed to smoke in a vehicle. Further, amongst the females in this study, those who are exposed to smoke in a vehicle suffer from less severe asthma

than those not exposed to smoke in a vehicle. For those in the study not exposed to smoke in a vehicle, asthma severity is marginally worse for the males.

Interpretation of the 'fear * breakfast habits' interaction showed that, compared to those who do not experience fear and eat breakfast daily, there is a significant chance that those who do experience fear but do not eat breakfast daily will suffer from worse asthma. Results also indicate that for those who eat breakfast daily, worse asthma is experienced by those who do not experience fear than by those who do experience fear. Furthermore, for those who do not experience fear, children who eat breakfast daily have marginally worse asthma than those who don't eat breakfast daily. Whereas with s-CA the classifications as supplementary variables of those variables included in the interactions enabled the study of their positions relative to the asthma severity categories (male children suffer from worse asthma than female children [30, 31]), this was not possible with the multiple imputation approach.

While on the surface these methods produce the same overall results, a deeper study of the results identified several differences in the outcomes from these methods.

Whereas with multiple imputation and ordinal regression the interpretation of results indicated the relative severity of asthma from one category to another category of a specific variable, with the application of CA we were able to identify factors associated with the specific asthma severity classifications. To illustrate this point, analysis with multiple imputation and ordinal regression showed that worse asthma is experienced by those who had neonatal care than by those who did not have any neonatal care. On the other hand, results from s-CA were more specific and having had neonatal care

was shown to be associated with moderate to severe asthma while not having neonatal care is associated with mild intermittent or no asthma.

Compared to the multiple imputation approach, CA was also able to identify specific associations that distinguished between the different levels of variables. This can be seen with the 'TV watching' variable. Results from the multiple imputation approach indicate that the amount of TV watched is inversely proportional to the severity of the asthma. With s-CA, by examining the

positions of these variable categories in the graphical display (Figure 1) as well as the decomposition of the inertia (Table 3), we see that watching 1 hour of TV a day (T1) is associated with moderate to severe asthma; watching between 1 and 3 hours a day (TV2) is associated with mild persistent asthma; and watching more than 3 hours a day (T3) is associated with mild intermittent or no asthma. Thus a finer distinction is possible regarding categories of variables and their associations with levels of asthma severity.

By using the graphical display produced by s-CA, it is possible to identify inter-variable relationships that do not include the asthma severity variable. For example, the positions of the variables I3, Np1, SD and VN indicate that they share some relationship. This is not possible with the MI approach.

With CA, it was also possible to compare the strengths of association with asthma severity of several predictor variables. For instance, from the positioning of the points on the display, we can deduce that while the risk of having moderate to severe asthma from smoke exposure in a vehicle exceeds the risk from smoke exposure in the

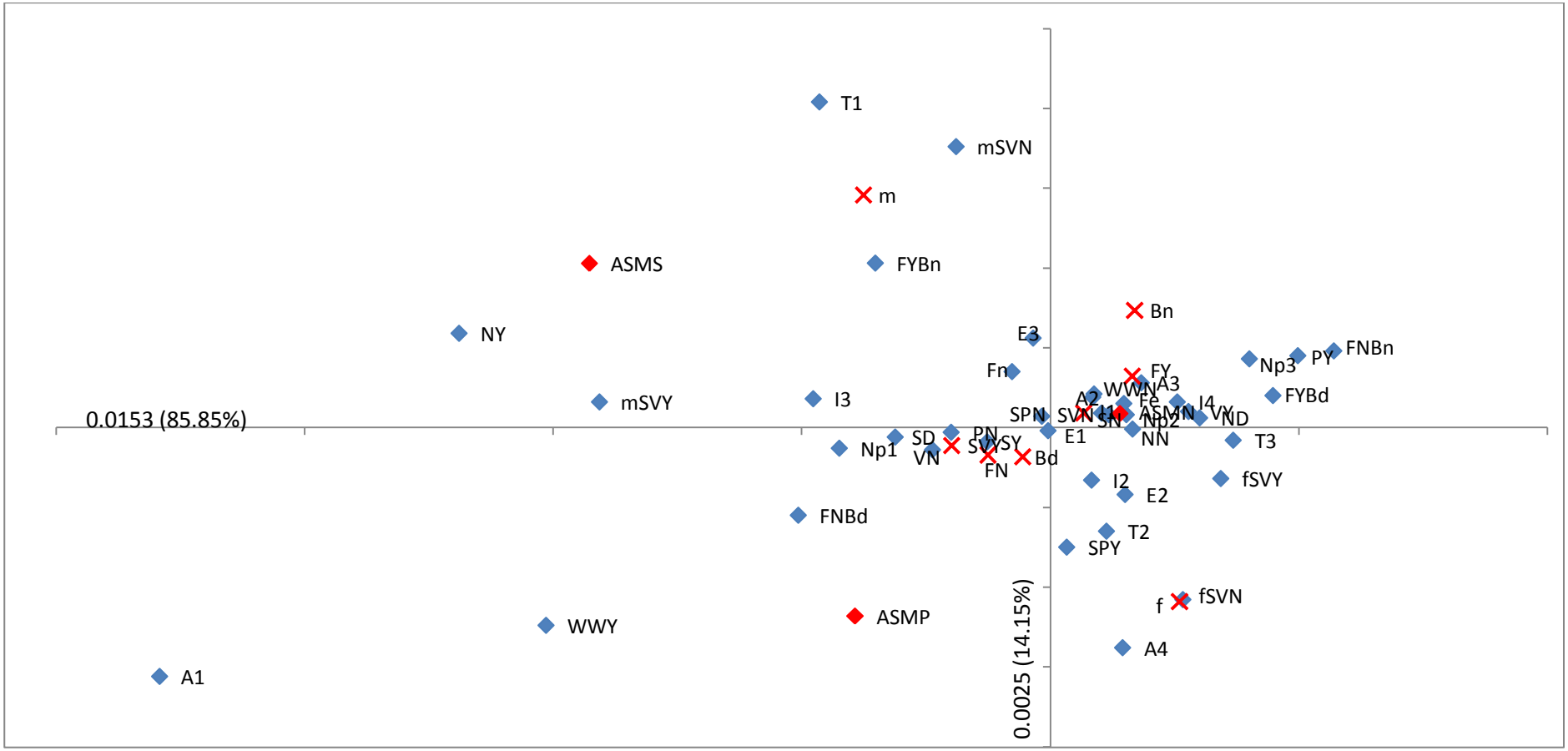



Figure 1: Subset CA map of a contingency tables with the row points represented by ◆ and the column points by ◆ projected onto the plane of the first and second principal axes. Supplementary points are represented with **X**. Values on the axes indicate principal inertias and their respective percentages of total inertia.

home [32] or smoke exposure in utero, the greatest risk is from air pollution as experienced in the South Durban region.

All these factors discussed above illustrate the extent of the usefulness of the graphical display produced by s-CA as a tool to identify inter-variable relationships.

Unlike the analysis with multiple imputation and ordinal regression, inter-variable relationships found to exist with the application of CA cannot be assumed to be statistically significant. While the relative strength of associations can be deduced by examining the angles that the points made with each other and with the principal axes in the graphical display [9], these results cannot be projected onto a broader population. Results from s-CA indicated that the association of ASMS (moderate to severe asthma) with NY (having had neonatal care) is stronger than its association with SD (South Durban) as seen by the relative size of the angles between them. These results are confirmed in the multiple imputation and ordinal regression analysis and, in addition, the significance of the association between neonatal care and asthma severity is indicated.

Because multiple imputation is computationally intensive, complications and limitations can be encountered. This can occur with large data sets and even more so when a large number of variables suffer from missingness [33, 34]. The need to include many interactions in the imputation model in order to ensure that it is more general than the analysis model, is often not feasible and computationally not possible [33] – especially with data sets that have a large number of variables. We did not encounter these problems with this analysis despite the seemingly large number of variables. In

fact, in a previous study using this data [8], 10 interactions were included with no problems being experienced. In contrast, computationally, CA can cope with large numbers of variables and interactions, but this can cause overcrowding in the display which makes it difficult to identify points and interpret relationships between them. It is for this reason that we limited the number of interactions in this study to two. The possibility does, however, exist with s-CA to include more interactions and analyse them as a separate subset.

Preliminary analysis of this data set indicated that the missingness is at best MAR with a possibility of some MNAR present [8]. Because multiple imputation produces unbiased estimates providing the missingness is at worst MAR, it was necessary to include, in the imputation model, variables associated with the missingness of the incomplete variables, the outcome variable – asthma severity - as well as the two interactions chosen for the analysis model. This inclusion of carefully selected variables should produce acceptable results even if some MNAR is present [35]. In contrast to this, CA and its variants are not constrained by complexities of models or distribution requirements. It is also not sensitive to the missingness mechanism in the data [36]. Therefore no special adjustments were needed to counteract the possibility of some MNAR missingness. The only adjustment needed in this study was to categorize the interval variable ‘age’. While non-negative categorical data is a requirement of CA, it is generally a straightforward exercise to achieve this condition.

The fact that only a few of the variables in the multiple imputation/ordinal regression analysis were significantly associated with asthma severity is consistent with the

results from s-CA. The visible bunching up of the points in the graphical display and the low inertia values – a total of only 0.0178 - indicate that only a limited amount of variability is present in this data [37].

## Conclusion

Non-response is a reality in survey data and needs to be handled appropriately. We have demonstrated the use of multiple imputation in conjunction with ordinal regression as well as CA as applied to the subset of measured data to analyse categorical data that suffer from missingness. We have also illustrated how interactions can be added to an analysis with s-CA. We found that general relationships between the environmental, socio-economic, genetic and behavioural variables and asthma severity were consistent across methods. Each method offers a different set of advantages in their applications. Analysis with s-CA is less demanding than with the multiple imputation approach – both in terms of conditions and the computational process – and finer distinctions in the inter-variable relationships can be made. These relationships are, however, 'looser' than those obtained from the multiple imputation approach and significance cannot be claimed. Despite their differences, the results produced in this study provide support for the greater use of less restrictive and less computationally intensive graphical methods to analyse categorical data that suffer from missingness.

## Competing interests

The authors have no competing interests to declare.

## Authors' contributions

GH analysed and interpreted the data, drafted the original article and made suggested revisions. RN was involved in the collection of the data, supervised the study, critically reviewed the article and approved the final version to be published. TZ supervised the study, critically reviewed the article and approved the final version to be published and DN supervised the study and approved the final version to be published.

## Acknowledgments

This work was supported by eThekwini Metropolitan Municipality (local government) – Contract No 1A-103; Medical Research Council of South Africa and University of KwaZulu-Natal – Research Funds. We acknowledge Dr Graciella Mentz for her part in the earlier stages of the project with the data collection and cleaning.

## References

1. Little RJA, Rubin DB: **Statistical Analysis With Missing Data**. New York: Wiley; 1987.
2. Eekhout I, de Boer RM, Twisk JW, de Vet HC, Heymans MW: **Missing Data: A Systematic Review of How They Are Reported and Handled**. *Epidemiology* 2012, **23**(5):729-732.
3. Greenland S, Finkle WD: **A critical look at methods for handling missing covariates in epidemiologic regression analyses**. *American Journal of Epidemiology* 1995, **142**(12):1255-1264.
4. Scheuren F: **Multiple imputation: How it began and continues**. *The American Statistician* 2005, **59**(4):315-319.
5. Schafer J: **NORM: Multiple imputation of incomplete multivariate data under a normal model [Computer software]**. In. University Park: Pennsylvania State University, Department of Statistics; 1999.
6. Van Buuren S, Boshuizen HC, Knook DL: **Multiple imputation of missing blood pressure covariates in survival analysis**. *Statistics in Medicine* 1999, **18**(6):681-694.
7. Greenacre M, Pardo R: **Subset correspondence analysis visualizing relationships among a selected set of response categories from a questionnaire survey**. *Sociological Methods & Research* 2006, **35**(2):193-218.
8. Hendry GM, Naidoo RN, Zewotir T, North D, Mentz G: **Model development including interactions with multiple imputed data**. *BMC Medical Research Methodology* 2014, **14**(1):1-11.

9. Hendry G, North D, Zewotir T, Naidoo R: **The application of subset correspondence analysis to address the problem of missing data in a study on asthma severity in childhood**. *Statistics in Medicine* 2014, **33**(22):3882-3893.
10. Rubin DB: **Multiple imputation for nonresponse in surveys**, vol. 81. New York: John Wiley & Sons; 2004.
11. Azur MJ, Stuart EA, Frangakis C, Leaf PJ: **Multiple imputation by chained equations: what is it and how does it work?** *International Journal of Methods in Psychiatric Research* 2011, **20**(1):40-49.
12. Greenacre MJ: **Theory And Applications Of Correspondence Analysis**. London: Academic Press; 1984.
13. Graham JW: **Missing data: Analysis and design**. New York: Springer; 2012.
14. Raghunathan TE, Solenberger PW, Van Hoewyk J: **IVEware: Imputation and variance estimation software**. In: *Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan.* Ann Arbor, MI; 2002.
15. Rubin DB: **Multiple imputation for nonresponse in surveys**. New York: Wiley 2004.
16. White IR, Royston P, Wood AM: **Multiple imputation using chained equations: issues and guidance for practice**. *Statistics in Medicine* 2011, **30**(4):377-399.
17. Wood AM, White IR, Royston P: **How should variable selection be performed with multiply imputed data?** *Statistics in Medicine* 2008, **27**(17):3227-3246.
18. Torres-Lacomba A: **Correspondence analysis and categorical conjoint measurement**. In: *Multiple Correspondence Analysis and Related Methods.* edn. Edited by Greenacre MJ, Blasius J. Boca Raton: Chapman & Hall/CRC; 2006: 421 - 432.
19. Asher MI, Montefort S, Björkstén B, Lai C, Strachan DP, Weiland SK, Williams H: **ISAAC Phase Three Study Group: Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys**. *Lancet* 2006, **368**(9537):733-743.
20. Mai XM, Gäddlin PO, Nilsson L, Finnström O, Björkstén B, Jenmalm MC, Leijon I: **Asthma, lung function and allergy in 12-year-old children with very low birth weight: A prospective study**. *Pediatric Allergy and Immunology* 2003, **14**(3):184-192.
21. Charoenca N, Kungskulniti N, Tipayamongkholgul M, Sujirarat D, Lohchindarat S, Mock J, Hamann SL: **Determining the burden of secondhand smoke exposure on the respiratory health of Thai children**. *Tobacco Induced Diseases* 2013, **11**(1):7-12.
22. Ehrlich R, Kattan M, Godbold J, Saltzberg DS, Grimm KT, Landrigan P, Lilienfeld D: **Childhood asthma and passive smoking**. *American Review of Respiratory Diseases* 1992, **145**(3):594-599.
23. Sendzik T, Fong GT, Travers MJ, Hyland A: **An experimental investigation of tobacco smoke pollution in cars**. *Nicotine & Tobacco Research* 2009, **11**(6):627-634.
24. DiFranza JR, Aligne CA, Weitzman M: **Prenatal and postnatal environmental tobacco smoke exposure and children's health**. *Pediatrics* 2004, **113**(Supplement 3):1007-1015.
25. Neidell MJ: **Air pollution, health, and socio-economic status: the effect of outdoor air quality on childhood asthma**. *Journal of health economics* 2004, **23**(6):1209-1236.
26. Peden DB: **The epidemiology and genetics of asthma risk associated with air pollution**. *Journal of Allergy and Clinical Immunology* 2005, **115**(2):213-219.
27. Jarvis D, Chinn S, Luczynska C, Burney P: **The association of family size with atopy and atopic disease**. *Clinical & Experimental Allergy* 1997, **27**(3):240-245.
28. Becher R, Hongslo JK, Jantunen MJ, Dybing E: **Environmental chemicals relevant for respiratory hypersensitivity: the indoor environment**. *Toxicology letters* 1996, **86**(2-3):155-162.

29. Venables KM, Chan-Yeung M: **Occupational asthma**. *The Lancet* 1997, **349**(9063):1465-1469.
30. Almqvist C, Worm M, Leynaert B: **Impact of gender on asthma in childhood and adolescence: a GA2LEN review**. *Allergy* 2007, **63**(1):47-57.
31. Bonner J: **The epidemiology and natural history of asthma**. *Clinics in Chest Medicine* 1984, **5**(4):557-565.
32. Sly PD, Deverell M, Kusel MM, Holt PG: **Exposure to environmental tobacco smoke in cars increases the risk of persistent wheeze in adolescents**. *Medical Journal of Australia* 2007, **186**(6):322-322.
33. Lee KJ, Carlin JB: **Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation**. *American journal of epidemiology* 2010, **171**(5):624-632.
34. Van Buuren S: **Multiple imputation of discrete and continuous data by fully conditional specification**. *Statistical Methods in Medical Research* 2007, **16**(3):219-242.
35. Graham JW, Hofer SM, Donaldson SI, MacKinnon DP, Schafer JL: **Analysis with missing data in prevention research**. *The Science of Prevention: Methodological advances from alcohol and substance abuse research* 1997, **1**:325-366.
36. Hendry GM, Zewotir T, Naidoo RN, North D: **The Effect of the Mechanism and Amount of Missingness on Subset Correspondence Analysis**. *Correspondence in Statistics* 2015, **Under review**.
37. Greenacre M: **Correspondence analysis in medical research**. *Statistical Methods in Medical Research* 1992, **1**(1):97-117.