A COMPARATIVE STUDY OF MULTIPLE IMPUTATION AND SUBSET CORRESPONDENCE ANALYSIS IN DEALING WITH MISSING DATA

G. M. Hendry [1]
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville Campus, University Road, Westville, Durban, South Africa
e-mail: *hendryfam@telkomsa.net*

T. Zewotir
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville Campus, University Road, Westville, Durban, South Africa
e-mail: *Zewotir@ukzn.ac.za*

R. N. Naidoo
Discipline of Occupational and Environmental Health, School of Nursing and Public Health, University of KwaZulu-Natal, Durban, South Africa
e-mail: *naidoon@ukzn.ac.za*

D. North
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville Campus, University Road, Westville, Durban, South Africa
e-mail: *northd@ukzn.ac.za*

Key words: Missing data, Multiple imputation, Subset correspondence analysis.

Abstract:
Methods: 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.

[1] Corresponding author.

AMS: 62-07, 62H25, 62J99, 62P12

1. 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 (1987). 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 (Eekhout et al., 2012), 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 (Little and Rubin, 1987; Greenland and Finkle, 1995). 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 (Scheuren, 2005). 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 (non-monotonic missingness) are: multiple imputation based on the multivariate normal distribution (MVNI), available in a standalone package - NORM - developed by Schafer (1999); and an algorithm known as "fully conditional specification" (FCS) or "chained equations" - implemented by, amongst others, Van Buuren, Boshuizen and Knook (1999) - 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 types.

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 (2006). 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 (Hendry, Naidoo, Zewotir, North and Mentz, 2014b; Hendry, North, Zewotir and Naidoo, 2014a), 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.

2. 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.

Table 1: Categories, code names and frequencies for all variables.

Variables	**Categories (code names) – count** ($N = 382$)								**Non-response – count (%)**
Gender	male (m)	163	female (f)	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 (FY)	165	no (FN)	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

2.1. 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 (Hendry et al., 2014b; Hendry et al., 2014a).

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 (Rubin, 2004).

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, Stuart, Frangakis and Leaf (2011).

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 (Greenacre, 1984). 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. Applying this variant of CA, a subset of the data matrix can be selected for analysis. CA is then applied to the subset with the important modification that the frequencies of variable categories relative to each other are calculated for the full matrix and retained in the analysis of the subset. A brief description of s-CA, as applied to a contingency table $\boldsymbol{N}$, follows.

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 $\mathbf{r}$ and $\mathbf{c}$, as marginal densities. The elements of $\mathbf{r}$ and $\mathbf{c}$, termed masses, are a measure of the relative importance of each row and column point. Vectors of relative frequencies, called row (column) profiles, are formed by dividing each element of a row (column) by its respective row (column) sum. These profiles define the two clouds of points, one for rows and one for columns, in multi-dimensional weighted Euclidean space.

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 normalize the correspondence matrix and results in a matrix of standardised residuals, $\boldsymbol{S}$. 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 the CA process is "interrupted" to implement the "adjustment" needed for s-CA. From the matrix $\boldsymbol{S}$ 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 $\boldsymbol{S}^*$. It is important to note that marginal densities, $\mathbf{r}$ and $\mathbf{c}$, for the full matrix are retained for all future calculations (Greenacre and Pardo, 2006).

By performing a singular value decomposition (SVD) on $\boldsymbol{S}^*$, the principal coordinates of the points, i.e. the coordinates with respect to the principal axes are defined. It is these coordinates that are used to produce the graphical display of the row points and the column points in a joint space.

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

2.2. 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 of missingness.

In the multiple imputation approach, based on the amount of missingness present, 20 data sets were imputed (Graham, 2012). To ensure stability of the parameters, ten iterations of the imputation process were completed between each retained complete data set (Raghunathan, Solenberger and Hoewyk, 2002).

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 (Rubin, 2004). 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 (Graham, 2012).

The identification of interactions, in the presence of missing data, presents a challenge (White, Royston and Wood, 1997; White, Royston and Wood, 2008). In a previous study using this data set, this problem was addressed and 10 interactions were identified as being significant (Hendry et al., 2014b) . 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 continuous (age) and 16 categorical - were treated as predictor variables.

On the other 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 (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.

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.

3. 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 (Asher, Montefort, Bjorksten, Lai, Strachan, Weiland and Williams, 2006); have had some special neonatal care* (Mai, Gaddlin, Nilsson, Finnstrom, Bjorksten, Jenmalm and Leijon, 2003); are exposed to smoke in the home* (Charoenca, Kungskulniti, Tipayamongkholgul, Sujirarat, Lohchindarat, Mock and Hamann, 2013; Ehrlich, Kattan, Godbold, Saltzberg, Grimm, Landrigan and Lilienfeld, 1992), in vehicles (Sendzik, T.Fong, Travers and Hyland, 2009), in utero (DiFranza, Aligne and Weitzman, 2004) and in the form of air pollution (Neidell, 2004; Peden, 2005); lived in a home with up to 4 people* (Jarvis, Chinn, Luczynska and Burney, 1997); come from a R4501 - R10000 income household; do not always have enough food; are exposed to low concentrations of compounds and pollutants (Becher, Hongslo, Jantunen and Dybing, 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 was explored for reasons for these anomalies, none

Table 2: Estimated Coefficients (EST) and Standard Errors (SE).

Predictor	Reference Category	Category	FCS($N = 382$)	
			EST	SE
Gender	Female	Male	0.039	0.36
Neonatal care	No	Yes	0.847*	0.39
Fear	No	Yes	-1.042*	0.41
Smoked while pregnant	No	Yes	0.379	0.49
Smokers in home	No	Yes	0.701*	0.31
Smoke in vehicles	No	Yes	-0.706	0.51
Exercise	>4 times a week	Up to once a week	0.044	0.38
		2 – 4 times a week	0.011	0.38
TV watching	>3 hours a day	Up to 1 hour a day	0.786	0.47
		1 – 3 hours a day	0.046	0.43
Number people in home	8+	1 - 4	0.981*	0.48
		4 - 7	0.381	0.49
Income	R100001+	up to R1000	-0.133	0.61
		R1001 – R4500	0.017	0.56
		R4501 – R10000	0.697	0.52
Food availability	Enough	Not always enough	0.756	0.41
Work'nWear	No	Yes	0.402	0.46
Pets ever	No	Yes	-1.072*	0.4
Area	North Durban	South Durban	0.595	0.31
Breakfast habits	Daily	Not daily	-1.011*	0.49
Violence	No	Yes	-0.709*	0.34
Age			-0.247	0.16
Fear * Breakfast	No/Daily	Yes/Not daily	2.338*	0.73
Gender * SmokeVehicle	Female/No	Male/Yes	1.811*	0.7

ND – North Durban; SD – South Durban; preg – pregnant;
FCS - Multiple imputed FCS
*Significant at the 0.05 level

were found. We 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. 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 do not 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, Worm and Leynaert, 2007; Bonner, 1984)), 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.), which is an exact representation of the points in 2-dimensional space, 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

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
NY	9	3	-476	985	129	59	15	12
NN	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
Np1	22	18	-170	994	41	-13	6	1
Np2	27	7	61	983	7	8	17	1
Np3	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
SD	33	15	-125	997	33	-6	3	1
ND	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; Coordinates ($k = \ldots$); 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
FNBn	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
ASPN	806	781	56	975	168	9	25	26
SUPPLEMENTARY								
male			-150	516		146	484	
fem			107	494		-108	506	
FY			66	808		32	192	
FN			-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; Coordinates ($k = \ldots$); contributions of axes to the inertia of the variables (COR); and contributions of variables to the inertia of the axes (CTR)*.

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 home (Sly, Deverell, Kusel and Holt, 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 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 (Hendry et al., 2014a) , 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.

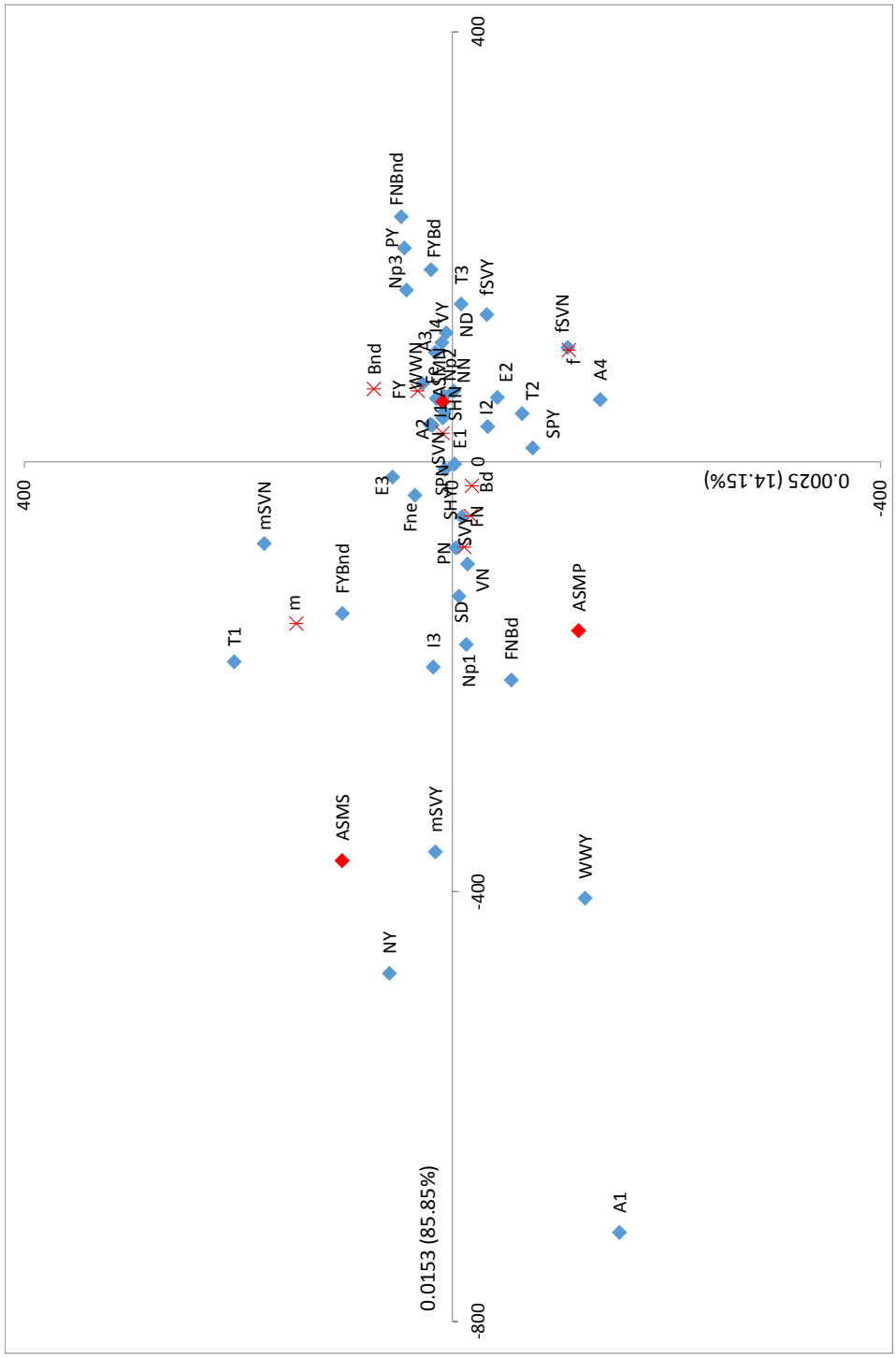

Figure 1: s-CA map of a contingency table with the row points represented by ◆ and column points by ◆ represented exactly in the plane of the first and second principal axes. Supplementary points are represented with ✳. Values on the axes indicate principal inertias and their respective percentages of total inertia.

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 (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. 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 (Hendry et al., 2014b), 10 interactions were included with no problems 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 (Hendry et al., 2014b). 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 (Graham, Hofer, Donaldson, MacKinnon and Schafer, 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, Zewotir, Naidoo and North, 2016). 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 (Greenacre, 1992).

While we chose to use a symmetrical map to graphically display both row and column points in the same space, given that this data is exactly represented in two dimensions, another possibility is to construct a ternary plot.

4. 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. When applied to this

data set, we found that general relationships between the environmental, socio-economic, genetic and behavioural variables and asthma severity were consistent across methods. Further investigations should be made to test the methods with other data sets. 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. It is suggested that it could be useful to apply s-CA in the exploratory stages of a research project before using a model-based method.

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 acknowledge Dr Graciella Mentz for her part in the earlier stages of the project with the data collection and cleaning. The authors are grateful to the anonymous reviewers for their helpful comments.

References

ALMQVIST, C., WORM, M., AND LEYNAERT, M. (2007). Impact of gender on asthma in childhood and adolescence: a GA2LEN review. *Allergy*, **63** (1), 47–57.

ASHER, M. I., MONTEFORT, S., BJORKSTEN, B., LAI, C., STRACHAN, D. P., WEILAND, S. K., AND WILLIAMS, H. (2006). 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*, **368**, 733 – 743.

AZUR, M. J., STUART, E. A., FRANGAKIS, C., AND 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** (1), 40–49.

BECHER, R., HONGSLO, J. K., JANTUNEN, M. J., AND DYBING, E. (1996). Environmental chemicals relevant for respiratory hypersensitivity: the indoor environment. *Toxicology letters*, **86** (2-3), 155–156.

BONNER, J. (1984). The epidemiology and natural history of asthma. *Clinics in Chest Medicine*, **5** (4), 557–565.

CHAROENCA, N., KUNGSKULNITI, N., TIPAYAMONGKHOLGUL, M., SUJIRARAT, D., LOHCHINDARAT, S., MOCK, J., AND HAMANN, S. L. (2013). Determining the burden of secondhand smoke exposure on the respiratory health of Thai children. *Tobacco Induced Diseases*, **11** (1), 7–12.

DiFranza, J. R., Aligne, C. A., and Weitzman, M. (2004). Prenatal and postnatal environmental tobacco smoke exposure and childrenŠs health. *Pediatrics*, **113** (Supplement 3), 1007–1015.

Eekhout, I., de Boer, R. M., Twisk, J. W., de Vet, H. C., and Heymans, M. W. (2012). Missing data: A systematic review of how they are reported and handled. *Epidemiology*, **23** (5), 729–732.

Ehrlich, R., Kattan, M., Godbold, J., Saltzberg, D. S., Grimm, K. T., Landrigan, P., and Lilienfeld, D. (1992). Childhood asthma and passive smoking. *American Review of Respiratory Diseases*, **145** (3), 594–599.

Graham, J. W. (2012). *Missing Data: Analysis and Design*. Springer: New York.

Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., and 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. and Pardo, R. (2006). Subset correspondence analysis visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods and Research*, **35** (2), 193–218.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press: London.

Greenacre, M. J. (1992). Correspondence analysis in medical research. *Statistical Methods in Medical Research*, **1** (1), 97–117.

Greenland, S. and 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., and 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** (22), 3882–3893.

Hendry, G. M., Naidoo, R. N., Zewotir, T., North, D., and Mentz, G. (2014b). Model development including interactions with multiple imputed data. *BMC Medical Research Methodology*, **14** (1), 1–11.

Hendry, G. M., Zewotir, T., Naidoo, R. N., and North, D. (2016). The effect of the mechanism and amount of missingness on subset correspondence analysis. *Communications in Statistics – Simulation and Computation*, **In Press**. DOI: 10.1080/03610918.2016.1224349.

Jarvis, D., Chinn, S., Luczynska, C., and Burney, P. (1997). The association of family size with atopy and atopic disease. *Clinical and Experimental Allergy*, **27** (3), 240–245.

Lee, K. J. and Carlin, J. B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*, **171** (5), 624–632.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis With Missing Data*. Wiley: New York.

Mai, X. M., Gaddlin, P. O., Nilsson, L., Finnstrom, O., Bjorksten, B., Jenmalm, M. C., and 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** (3), 184–192.

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** (6), 1209–1236.

PEDEN, D. B. (2005). The epidemiology and genetics of asthma risk associated with air pollution. *Journal of Allergy and Clinical Immunology*, **115** (2), 213–219.

RAGHUNATHAN, T. E., SOLENBERGER, P. W., AND HOEWYK, J. V. (2002). Iveware: Imputation and variance estimation software. *In Survey Methodology Program*. Ann Arbor: MI, Survey Research Center, Institute for Social Research, University of Michigan.

RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York.

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

SCHEUREN, F. (2005). Multiple imputation: How it began and continues. *The American Statistician*, **59** (4), 315–319.

SENDZIK, T., T.FONG, G., TRAVERS, M. J., AND HYLAND, A. (2009). An experimental investigation of tobacco smoke pollution in cars. *Nicotine and Tobacco Research*, **11** (6), 627–634.

SLY, P. D., DEVERELL, M., KUSEL, M. M., AND HOLT, P. G. (2007). Exposure to environmental tobacco smoke in cars increases the risk of persistent wheeze in adolescents. *Medical Journal of Australia*, **186** (6), 322–322.

TORRES-LACOMBA, A. (2006). Correspondence analysis and categorical conjoint measurement. *In* GREENACRE, M. J. AND BLASIUS, J. (Editors) *Multiple Correspondence Analysis and Related Methods*, 2nd edition. Chapman and Hall/CRC: Boca Raton, pp. 421–432.

VAN BUUREN, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, **16** (3), 219–242.

VAN BUUREN, S., BOSHUIZEN, H. C., AND KNOOK, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, **18** (6), 681–694.

VENABLES, K. M. AND CHAN-YEUNG, M. (1997). Occupational asthma. *The Lancet*, **349** (9063), 1465–1469.

WHITE, I. R., ROYSTON, P., AND WOOD, A. M. (1997). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, **30** (4), 377–399.

WHITE, I. R., ROYSTON, P., AND WOOD, A. M. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, **27** (17), 3227–3246.

Manuscript received, 2016-04-21, revised, 2016-10-11, accepted, 2016-10-29.