Imputing age at death for the deceased using household relationships.

By

Farai S. Chinanayi

Dissertation submitted in partial fulfilment of the Master of Philosophy Degree

Centre for Actuarial Research

Faculty of Commerce

University of Cape Town

May 2011

Plagiarism Declaration

I Farai S. Chinanayi know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own. I have used the Harvard convention for citation and referencing. Each contribution to, and quotation in, this project from the work(s) of other people has been attributed, and has been cited and referenced.

This project is my own work, neither whole or any part of it has been submitted for another degree at this institution or any other institution.

Date: .....

Signature: .....

Acknowledgements

I am greatly indebted to my supervisor Professor Rob Dorrington whom despite my ups and downs remained consistent in encouraging me to produce work. It was an honour to work with individuals from DataFirst and SALDRU whom assisted me with Statistics South Africa's dataset and NIDS dataset respectively, you are acknowledged. I also thank the Research Commons team for their word processing assistance and for making the library a haven of study.

The Center for Actuarial Research (CARe) personnel is acknowledged, not to forget to mention the course convener Associate Professor Tom Moultrie.

Last but not least I thank the Hewlet Foundation for funding my studies and the Beit trust for paying some of my living expenses in Cape Town.

Even though it may not be much, I dedicate this dissertation to the late industrious sister **Patience Hatinawedu Chinanayi (1 April 1974 to 27 October 2009)**

Farai S. Chinanayi
Cape Town, May 2011

Abstract

Immeasurable effort has been dedicated to estimating mortality using direct and indirect demographic techniques. However, literature available on methods applied to replacing missing values for non-responses in surveys or censuses so that these methods are implemented using sound data is sparse. The National Income and Dynamics Study (NIDS) household dataset includes the relationship of the deceased to the head of household variable. The relationship of the deceased to the head of household and the age of the head of household are incorporated into the Multiple Imputation (MI) technique proposed by Rubin (1987) to impute the missing ages at death for the deceased.

Current methods that are implemented to impute ages at death, such as Statistics South Africa's hotdeck, simple apportionment and Stata hotdeck are compared to the multiple imputation method. The comparisons are conducted by testing the differences between the imputed data and the data with ages, empirical density estimates, Kolmogorov–Smirnov tests and bivariate scatter plots.

The results are that the multiple imputation method shows similar age at death distributions for imputed data and observed data at relationship level, however, when the data are combined to form a complete dataset, statistically significant differences between the complete dataset and the observed ages arise. Even so, under the assumption that data are missing at random, such anomalies are permitted because the missing data may be a function of another variable observed in the dataset, in particular the relationship to the head of household. Statistics South Africa's hotdeck and the Stata hotdeck produce distributions of the complete dataset similar to the distribution of the observed ages. Notably, the apportionment method has its frequency distribution transformed proportionally and therefore it maintains the distribution of the observed data when the ages are allocated to the whole age range.

For external comparisons the proportion of deaths by age from the vital registration is used. Slight differences in age at death distributions are observed for all the imputation procedures. However, both hotdeck methods produce distributions

that peak in the middle ages, as noted by other researchers. All methods have lower proportions at older ages as compared to the vital registration.

Table of Contents

List of Figures

List of tables

Introduction

1.1 Background of the research

Data for surveys or censuses are collected, collated and processed into information, however, during the process data might be erroneous or unavailable for certain fields or records, therefore compromising the data quality (Manzari, 2004; Sande, 1982; Lepkowski *et al.*, 1987). Imputations and weights are applied to the data to compensate for these missing or erroneous data (Little, 1988; Rubin, 1987). These inconsistencies might be as a result of non-response, incorrect values or errors introduced in the later phase of processing. However, the major contributor to the inconsistencies in surveys is non-response, which might be in the form of item non-response or a unit non-response. This is when a value in a record is missing or the record for a respondent is missing. These non-responses are both solved by imputing the missing values in the relevant fields or weighting to attain the expected number of observed units (Rao, 1996).

Age at death data of the deceased in the past year as reported by households are also affected by these inconsistencies. The data are usually missing or occasionally implausible for many of the surveys or censuses because it is often difficult for respondents to remember such information especially if the deceased is not related to the respondent. Moultrie and Dorrington (2009) found that with regards to National Income Dynamics Study (NIDS) data, nearly 20 per cent of the recorded deaths(Royston, 2007; Paul, 2007) in the past twenty four months prior to the survey date had no age at death recorded. The metadata file for the 2001 census reveals that age at death data needed to be imputed in 7.75 per cent of the cases (Statistics South Africa, 2003a). Non-response to the age at death question of the deceased in a household is one of the questions that has a high rate of non-response (Statistics South Africa, 2003b).

Methods of editing and imputation of missing responses have been considered and developed by researchers. The most notable being the hotdeck method (Rockwell, 1975; United Nations, 2001). Hotdeck methods are widely used, for instance, Statistics South Africa implemented logical imputations and hotdecks (dynamic imputations) where it was not possible to impute logically for the 2001 census (Statistics South Africa, 2003a). Logical imputations are preferred, as they have better chances of producing accurate values (Charlton, 2004; Nordholt, 1998). However, logical imputations can only be used for those data items

with overlapping information, for instance, the race of a minor at a household, can be imputed from the race of the head of the household (Statistics South Africa, 2003b). On the other hand, hotdecks are sensitive to the size of the deck (Statistics South Africa, 2003a). If the deck has few covariates it either increases repetition of values imputed or may impute more random values. That is imputation of repeated values and imputation of more random values occurs when data are missing in consecutive records and when data are missing sporadically, respectively. Conversely, if the deck has a large number of covariates it takes time to be updated and this may also lead to repetition of imputed values.

Dorrington *et al* (2004) suggest that Statistics South Africa for the census of 2001 should have define better covariates for the decks as heterogeneity of hotdeck responses needs to be minimised. The hotdeck imputation technique used by Statistics South Africa for the age at death of the deceased appears to distort the distribution of deaths by age, exaggerating number deaths for those aged below 40 years (Dorrington *et al.*, 2004). Williams (1998) reached similar conclusions on the 1990 United States of America census data namely that hotdecking resulted in a peaked age distribution because covariates resemble groups, and therefore hotdecking results in the loss of information. Furthermore, he also found that some of the multi-variable characteristics noticeable in the age data are distorted when hotdecks were employed, but, the model based imputation that he used preserved these characteristics.

Alternatively, in the absence of any other information to assist in replacing missing data demographers frequently apportion those with missing ages either according to the whole age range or to adult ages, above 20 years on the assumption that respondents would know children's ages. Another method that has been explored is cold deck imputation, which entails imputing values from a survey conducted in the past. This is especially applicable to panel surveys, where, because of attrition, which a respondent responds to the initial survey but fails to respond in the subsequent surveys (Lepkowski *et al.*, 1987).

Multiple imputation methods proposed in Rubin (1987) have been extensively implemented in health economics and epidemiology but not particularly to demographic data (Kmetic *et al.*, 2002; Faucett *et al.*, 2002).

1.2 Aims and objectives of the research

This research seeks to investigate if incorporating the relationship of the deceased to the head of the household variable in a method that replaces missing ages at death enhances

imputations of the ages at death. The investigation makes use of the National Income Dynamics Study (NIDS) wave 1 dataset to compare the multiple imputation procedure incorporating the relationship status to the methods that are currently in use, such as Statistics South Africa's hotdeck, Stata hotdeck, simple apportionment across the entire age range or limited to adult ages. The usefulness of the methods are assessed by comparing the resulting proportion of deaths by age against that from vital registration

1.3 Statement of the problem

The National Income and Dynamics Study (NIDS) asked, in addition to the usual census questions, about the relationship of the deceased to the head of household. This study analyses the NIDS data on the relationship of the deceased to the head of household and investigates if this relationship variable assists with the allocation of age at death for those deceased for whom age is missing. In the process of comparing current methods in use to the formulated method, the research interrogates and investigates the advantages of adopting a method which incorporates the relationship of the deceased to the head of the household.

1.4 Organisation of the dissertation

This research is presented in five chapters, the first being the Introduction. The literature review describing the relevant methods follows in Chapter 2. Chapter 3 explains the methods implemented to achieve the aims of the research. After describing the methods the results of implementing these methods are analysed and evaluated in Chapter 4. In Chapter 5, the discussions and the conclusions of the research are presented.

Chapter 2 Literature review

This chapter describes procedures undertaken in data processing with the intention of describing methods used to replace missing or implausible ages. External and internal techniques for investigating the quality of imputations are discussed. Ages at death data distributions for South Africa are also reviewed.

2.1 Types of missing values

Missing data take several forms. Data may be missing completely at random, missing at random or not missing at random (Scheffer, 2002; Peugh and Enders, 2004; Rubin, 1976). Data are missing completely at random if missing data do not depend on the variable with missing values or any other observed variable. The implications of this are that data are observed at random and also data are missing at random, meaning that data were collected randomly with the probability of missing being equal for all units. For cases that have variables with missing data to be deleted without violation of the underlying assumptions in data collection and analysis, data should have been observed and collected randomly (Scheffer, 2002). Rubin (1976) observes that the condition for a case deletion is difficult to satisfy, therefore it is rare that valid analyses could be made after deleting cases with missing observations. Furthermore, methods that are used on complete data may not be easily employed on data with non-response, consequently less efficient estimators are estimated because of the smaller sample size when cases are deleted (Rubin, 1987). When data are missing at random it means that the missing data may be dependent on other observed variables but not conditional on itself (Scheffer, 2002; Abayomi *et al.*, 2008). When data are not missing at random, it means that data are missing on condition of the missing value itself (Scheffer, 2002). This is the most difficult situation to impute or model, because the probability of a value missing varies and the missing observation cannot be explained by any other observed variable (Abayomi *et al.*, 2008).

Missing data in surveys are usually missing at random. Unfortunately, as Abayomi *et al.* (2008) point out missing mechanisms cannot be easily tested on observed data because the data needed to do this are usually unavailable, however, the models that are fitted to the data can be tested for goodness of fit. When data are missing completely at random or missing at random and the missing data percentage is not high, likelihood based imputations such as

Multiple Imputation (MI), Expectation Maximisation (EM) and regression may be successfully employed (Scheffer, 2002).

2.2 Statistical editing and imputation

Editing and imputation are terms used to imply two distinct statistical procedures, however, imputation may be part of editing depending on the method of imputation. For instance, logical or deductive imputations are also known as edits (Nordholt, 1998). In data processing, data are categorised into fields and records. Fields are single items that are expected to be completed on the questionnaire, such as the age of the respondent, interview date and other questions according to the questionnaire. Fields may contain qualitative or quantitative data. On the other hand, a record is a combination of fields pertaining to a single entity, such as a person or a household. A person's record may include fields such as the date of birth, the age and the place of birth for the respondent. For a household record, the fields may be the number of people in the household, the ethnicity of the household head, the income and expenditure of the household, for a survey or a census questionnaire. The household record may also include each person's record for those residing at the particular household.

Editing is the process of ensuring that data are free from errors through altering or correcting the data (United Nations, 2000). Non-sampling errors such as data entry errors, processing errors and interviewer errors may be detected and corrected. The process of editing can be manual or mechanical. The process of editing is conducted through edits which are defined by National Statistical Services (1997:10) as, "a logical condition or a restriction to the value of a datum or response which must be met if the datum is to be considered correct." Therefore the conditions that the data are to meet are applied and if the data fail to meet a condition they become failed edits and corrective actions are taken, although it is not always the case that corrective actions are taken.

Editing is a procedure conducted throughout the process of data preparation. There are four main types of edits, namely; validation edits, missing data edits, logical edits and consistency edits (National Statistical Services, 1997).Validation edits are checks that are done for every field of a record to ensure that they contain a valid value and that the entries are consistent with each other. For instance, checking that the entries for the same record make sense when considered together. Missing data edits check if relevant fields for a respondent were completed, for instance, an answer to a specific question may determine that the other

questions are to be answered. Logical edits correct for contradictions between fields, for instance, when two fields such as age and ever having given birth, in the same record, reveal that a nine year old girl is a mother. Consistency or reconciliation edits ensure that arithmetic relationships are adhered to, for example, that the total number of births reported is the sum of male births and female births reported.

Imputation is the process of replacing missing, inconsistent or implausible values in the fields identified during editing (United Nations, 2001). Imputation is conducted through changing one or more responses in the respondents' records to make the records consistent and complete. Imputation completes the process of data editing after the preliminary consistency checks and occasional respondent contacts.

The practice of imputation has the advantages of reducing the cost of editing and expediting data analysis by curtailing the number of follow-ups and making the procedure internal (Giles, 1988). In addition, the retention of complete data sets provides consistence in the handling of missing data for analysis. However, there is a danger that the imputed values are regarded as the actual responses by the users of the data if the imputed values are not flagged or the imputation method may result in biased data if an inappropriate method of imputation is implemented. Therefore, for imputation to yield plausible results it requires informed knowledge of the meaning and the origins of the missing values (Refaat, 2007).

There are three categories of imputation, namely deductive imputations, deterministic imputations and stochastic imputations (Charlton, 2004; Nordholt, 1998). According to Nordholt (1998) deductive imputations, also known as edits, are imputed values deduced from other fields of the same record which have been edited. For instance, a missing age of a respondent can be derived from a correct date of birth. Imputation is conducted after data editing because some inconsistent records may serve as a donor, that is giving values to other records with missing values and therefore propagate inconsistencies if not corrected. Deterministic imputations are values derived from responses to other questions of other respondents or when the mean of available quantitative data is calculated and imputed in the missing field. On the other hand, imputations may be determined stochastically through generating random values. These three groups encompass the available methods for data imputation.

Properties of successful imputation as outlined by the United Nations (2001) are that the record with imputations should closely resemble a consistent complete record, that is, a

record that has successfully gone through edits. The implications of this requirement are that the number of fields imputed for one record should be limited, as several imputations in one record tend to distort the relationships between fields of that record. If the numbers of imputations are few it enhances the quality of the imputed data because many of the imputation methods depend on the observed data. Furthermore, the record with imputed values should be consistent, that is, after imputation no values in different fields of the record should contradict any other values. Finally, both the imputed and the non-imputed records should be retained to allow researchers to evaluate the degree and effectiveness of the imputation, flagging all imputations so that the method and procedure of imputation are transparent (United Nations, 2001).

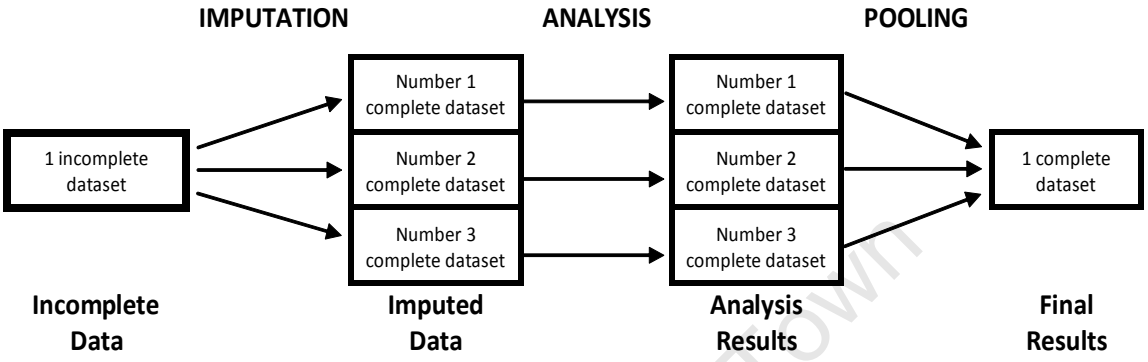
Desirable characteristics of successful imputation as proposed by Charlton (2004) are predictive accuracy, ranking accuracy, distributional accuracy, estimation accuracy and imputation plausibility. The five characteristics stated above are explained as follows; predictive accuracy is the resemblance of the imputed values to the estimated correct values, ranking accuracy is the order of the imputed values as they are required to be similar to the correct values, distributional accuracy is the similarity between the distributions of the imputed values and the correct values, estimation accuracy is the element that the imputed datasets should produce unbiased and efficient parameter inferences and imputation plausibility is the characteristic that the values imputed should be credible. The characteristics may not be independent or mutually exclusive, with, the type and the purpose of collecting the data determining the characteristic to be pursued (Charlton, 2004). For instance, when data are to be used for model building, the predictive accuracy and ranking accuracy are important (Charlton, 2004). On the other hand, if imputing ages at death, distributional accuracy and imputation plausibility are important.

2.3 Multiple imputation

Multiple imputation is a method of imputation, proposed by Rubin (1987), which imputes three or more values to each missing response of a variable on the basis of a regression of the non-missing values of the variable on one or more independent explanatory variables. Unique imputed values can then be estimated for each missing response, for example, the mean of the multiple imputations for that particular response. Figure 2.1 illustrates the imputation procedure showing three main stages that are imputation, analysis and pooling. The illustration

presents a case where three complete datasets are created from which the final one is constructed.

Figure 2.1 Multiple imputation procedure



Source: Derived from multiple imputation online (Buuren, 2005). (Available at http://web.inter.nl.net/users/S.van.Buuren/mi/)

The three or more sets of imputed values for the dependent variable are derived from the curves resulting from the coefficients samples from the joint distribution of the parameters produced by the regression of the non-missing data on the independent variables. For instance, a linear regression model in the form of $y = b_0 + b_1 x + \varepsilon$ where ε is an error term with mean 0 and variance σ^2, may be fitted to replace the missing values. The creation of a set of imputations for the missing values is done in two steps. First, the values of b_0^*, b_1^* and σ^{*2} are drawn randomly from the joint distributions of the regression parameters. These distributions may be approximated by an inverse of the chi-squared distribution for σ^{*2} and a bivariate normal distribution for extracting b_0^* and b_1^* given σ^{*2}. Second, for each of the cases with missing age at death the imputation procedure for case i would be $y_i^* = b_0^* + b_1^* x_i + \varepsilon_i^*$ where x_i is the i^{th} observation of variable x and ε_i^* is drawn from a normal distribution with mean 0 and variance σ^2 (Parker and Schenker, 2007).

The first step reflects the uncertainty of the model being fitted to a sample and the second step reflects the uncertainty of the model itself. The imputation procedure to become a multiple imputation the above two steps are repeated independently at least three times.

Rubin (1987) showed that one is unlikely to need more than 10 sets of imputed values to capture all the information, by evaluating the efficiency of the estimators for the datasets. He finds that the efficiency of an estimator from m imputations is a function of $(1+\frac{\gamma}{m})^{-1}$, where γ is the rate of missing information for the quantity being estimated, calculated as $\frac{r+2/(df+3)}{r+1}$, where r is the relative increase in variance due to non-response, calculated as $\frac{(1+m^{-1})B}{\bar{U}}$, where *df* is the degrees of freedom and where $\bar{U}$ and B are the within imputation variance and between imputation variance, respectively. Table 2.1 shows how the efficiency changes according to the number of imputations and the rate of missing information.

Table 2.1 Efficiency of an estimator according to the number of imputations m and the rate of missing data γ

	γ				
m	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Source: Rubin (1987:114)

According to Rubin (1987) multiple imputation has the following advantages. First, the efficiency of estimation is increased through the random extraction of values to be imputed because the randomization attempts to represent the uncertainty around the missing value, when a missing observation is replaced with a single value it seems to suggest that the true value is substituted. Second, by generating complete datasets of imputed values, from repeated random draws using one model it is possible to draw inferences about additional variability due to missing values under that model, or when more than one model is employed to create complete datasets, it allows one to study the sensitivity of inferences of various models used to impute. Third, multiple imputation may use a single imputation method such as hot decking to create one of the completed datasets.

Incorporating hotdecks into multiple imputation methods is done by matching observed complete cases and incomplete cases to replace missing values so that one complete

dataset is created. To create multiple complete dataset other methods are used. Such methods of imputation are called 'ignorable models' for non-response.

The disadvantages of multiple imputation relative to single imputation methods such as hotdecking, are that multiple imputations are tedious because of the creation of many data sets, the processing of data requires more space in memory and hard disc in comparison to single imputation methods and this may prolong the processing time and the analysis may also entail the implementation of complex imputation models (Parker and Schenker, 2007; Rubin and Schenker, 1986).

2.4 Cold Deck imputation

This imputation method entails tabulating static covariate values that assign values to missing values in records (United Nations, 2001). The covariates are characteristics such as sex, age and ethnicity that are combined and used to match the complete record with values for variables of interest or the 'donor record' and the record with a missing value in the variables of interest or the 'receiving record'. The value assigned is specific to a particular combination of covariate values and this value is not updated throughout the imputation procedure (United Nations, 2000). The values for the covariates are predetermined from prior surveys or in the absence of reliable data, and the tabulations with the covariates are created from current data with valid responses. Ratios or proportions may be calculated and then applied to impute the missing values (Nordholt, 1998).

The imputation method can be illustrated by a cold deck that imputes missing student's mark, by employing the school's group and the performance level of the student as covariates. Table 2.2 shows that given the group of the school and the performance of the student a mark can be determined. For instance, a student that is learning at a group C school and is above average, a missing mark can be imputed to 60, while a student that is learning at a group A school and is below average, the mark of 40 can be imputed if the mark is missing. The values remain static during the processing of records.

Table 2.2 Cold deck to determine the mark of a student given the group of the school and the overall performance of the student.

	Student performance		
School's Group	**Above Average**	**Average**	**Below Average**
A	80	60	40
B	70	50	30
C	60	40	20

Cold deck methods are part of deterministic methods (Nordholt, 1998). The major limitation of deterministic imputation is that the distribution of the values of the variable with imputed values tends to be too peaked and hence the variance to be too low. This is caused by imputing the best prediction at record level (Nordholt, 1998; Williams, 1998).

2.5 Hotdeck imputation method

The hotdeck method was developed by the U S Census Bureau, but some refinements have been added by different statistical institutions (United Nations, 2001). The method uses one or more variables, known as covariates, to impute the missing values for a particular variable. The responses of individuals with similar characteristics are used to determine the response for the variable to be imputed (Parker and Schenker, 2007). The hotdeck is continually updated by taking up new values from valid fields during the processing of records. In essence, the hotdeck uses a value stored in the hotdeck to replace a missing value by matching the covariates when it encounters a record with a missing value that should be replaced and in the case of encountering a complete record the hotdeck is updated by retaining the valid values for the specific covariate values. Therefore, donor records are the complete records and the incomplete records receive values.

There are a number of forms that hotdecking may take. Sequential hotdecking is where the imputed value is extracted from the preceding complete matching record. Hierarchical hotdeck imputation, an improvement to the sequential hotdeck, creates a large deck which is collapsed in a hierarchical fashion in the absence of a donor until a donor is found. This means that when a record with a missing value is is encountered and five variables

are being employed as covariates to match complete record, the covariates are reduced to four when the covariates of the complete records do not match with those of the incomplete record. This procedure is repeated until a matching complete record is found.

In addition, the hotdeck method can be randomized, for instance, the random hotdeck within classes which selects a record at random from the matched donors. When there is one class it becomes a random hotdeck overall method. The regular hotdeck imputation method, also known as the dynamic hotdeck, continually updates the values on the deck. In addition, there is the single donor rule or record matching whereby, several missing values for different variables in one record are simultaneously imputed from one complete record, through a random hotdeck to first match a record to impute the missing value of main variable (United Nations, 2001).

More commonly the hotdecking method does not substitute a value that is predetermined. For each deterministic method there is a stochastic counterpart, achieved by adding residuals to the deterministic method (United Nations, 2001). The stochastic counterpart has the advantage of maintaining the frequency structure of the data file or the variances and co-variances. In addition, stochastic methods are perceived to accommodate the variability in record responses, because, respondents can have similar responses on several fields but different responses in one of the fields. However, the method may create infeasible imputed values.

Several hotdeck method variations have been mentioned above, the only procedure defining the variation being the method of selecting the 'donor' record (Kaiser, 1983). The variations were developed to improve on particular aspects of the imputation method. For instance, it was observed that the sequential hotdeck performed better in keeping the covariance structure of the samples. However, the overall quality of the missing value estimates was lacking (Kaiser, 1983). The different hotdeck methods all suffer from not being able to maintain the sample covariance, however, the sequential hotdeck produces better results in comparison to other hotdecking methods.

2.5.1 Statistics South Africa's AMORTALITY Hotdeck

Statistics South Africa's census questionnaire for 2001 included the following questions on the deaths in the household in the previous year; whether there were any deaths, the number of deaths, the month of death of each death, the sex of each of the deceased, the age of the

deceased, whether the death was accidental and whether the death was during or within six weeks of a pregnancy (Statistics South Africa, 2003a). Logical imputations preceded hot decking imputation because it assigned more reliable and consistent values in comparison to the latter method (Statistics South Africa, 2003a:7). Therefore, edits are conducted before any imputations from the hot deck method are done so that the inconsistencies are not propagated.

The algorithm employed by Statistics South Africa for imputing the age of the deceased for whom age is missing entails a hot deck named AMORTALITY. The variables included in the hot deck are: MM-ALL, which returns any valid month; MM-2000, which returns a month from October to December; MM-2001, which returns a month from January to October; YYYY, which returns the year of death for the deceased that is 2000 or 2001; SEX, which returns the value 1 or 2 for male or female respectively; AGE-ALL which returns an age from 0 to 120 years; AGE-FERT, which returns a child bearing age from 12 to 50 years; and ACC, which returns a yes or no value for whether the death was accidental or whether the death was due to natural causes. The hot deck can be said to be dependent upon the population groups and the death occurrence number as both of these are the rows. The population group is assumed to be the same as that of the head of the household. The death occurrence number is also included in the hot deck from 1 to up to 5 in the order in which the deaths have been reported, however, some households have more than 5 deaths in the previous year. The additional deaths are additional deaths which are recorded in an extended file.

The hotdeck AMORTALITY is applied after the hotdeck APREGNANT, the latter imputes the pregnancy status of the deceased, that is, whether or not she was pregnant at the time of her death, and therefore it only applies to females between the ages of 15 years and 50 years. When death happened in the state of being pregnant the imputed ages would range from 12 years to 50 years, otherwise they would range from 0 years to 120 years. The hotdeck APREGNANT is structured as follows, the hotdeck has the death occurrence numbers as columns and the population groups as rows so that a value would be updated each time the death occurrence number and the population group is valid. The imputation flags numbered from zero to four were included to describe whether the type of imputation was a logical imputation to a blank field, logical imputation to missing value after an implausible value, a hotdeck imputation to a blank field, hotdeck imputation to missing value after an implausible value is deleted, respectively.

2.6 Diagnostics of imputations

The diagnostics after imputations are both external and internal (Abayomi *et al.*, 2008). External techniques are employed by comparing the observed data to the complete data including the imputed data. This is also known as the reasonability test. For instance, Abayomi*et al.*(2008) using the Environmental Sustainability Index data realised that after imputation of the missing values of the variable, 'BODWAT' which is a measure of the industrial and organic pollutants per available freshwater, became bimodal, different from the observed data that had a single mode. The differences suggest the need to investigate the underlying assumptions of the imputation model and the need to edit the observed data. In addition, as a method of diagnosis for an imputation procedure, the similarities between complete data and an external benchmark may also be employed. On the other hand, internal diagnostics include, investigating the predictive model employed to impute missing values to establish the goodness of fit to the observed data (Easterling, 1976).

The two-sample Kolmogorov-Smirnov test is one of the most useful and general non-parametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples (Kirkman, 1996). The major advantages of the Kolmogorov-Smirnov test in the context of this research are that it is non-parametric and it does not make assumptions about the distribution of the data. The test statistic tests whether the two datasets come from significantly different distributions. The empirical cumulative distribution function, F_n, for n independent and identically distributed observations X_i is defined as $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x}$ where $I_{X_i \leq x}$ is the indicator function which is equal to 1 if $X_i \leq x$ and equal to zero otherwise. Then the test statistic is $D_{n_1,n_2} = sup_x \, |F_{1,n_1}(x) - F_{2,n_2}(x)|$, where $F_{1,n_1}(x)$ and $F_{2,n_2}(x)$ are the empirical distributions of the first and second dataset respectively. The null hypothesis is rejected at level α if $\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{n_1,n_2} > K_\alpha$.

When the Kolmogorov-Smirnov test is conducted on imputed data dependent on observed data, one should prepare empirical density plots to allow visual inspection of the tests (Abayomi *et al.*, 2008). Bivariate scatter plots are also important in checking the internal consistency by plotting both imputed and observed observations against a continuous predictor. The method has been used by Abayomi*et al* (2008) in investigating the imputation

mechanism for the environment survival index. The bivariate scatter plots were employed in conjunction with empirical density plots and this enabled differences to be assessed based on internal data with respect to external knowledge. In this case it was realised that both are important in classifying imputations as problematic. Thus by evaluating the imputations as problematic there is the need to investigate where the problem arises, and thus the methods discussed above do not pin point where the problem exactly arises. This is because anomalies arise as a result of wrong assumptions about the missing mechanism, errors in the data or the imputation mechanism.

When an imputation model is employed residuals may be plotted and the non-random patterns in the observed dataset may be captured to enhance each imputation value, therefore developing the imputation model by fitting a Locally Weighted Regression and Smoothing Scatter (LOWESS) plot (Cleveland, 1979; Abayomi *et al.*, 2008). The enhancement on the imputation model is conducted by fitting a LOWESS curve to the residual differences between the observed data and the predicted data against the observed data and then the imputed values are enhanced by using the LOWESS curve as the correct residual function (Abayomi *et al.*, 2008).

Chapter 3 Data Sources and Method

This chapter commences by presenting and analysing the characteristics of the National Income Dynamics Study (NIDS) dataset as these characteristics may inform the methods employed to impute missing values and assist in explaining the results found in the next chapter. After this, the imputation methods are explained in order they have been applied, namely; Multiple Imputation (MI) technique, Statistics South Africa's hotdeck, STATA hotdeck and the apportionment method.

3.1 Evaluation of the Data sources

The National Income and Dynamics Study (NIDS) collected data on a question not usually asked in a census, that is, the relationship of the deceased to household head, in addition to the usual variables that are included under household mortality in survey and census questionnaires, such as age at death of the deceased, causes of death, month and year of death, the sex and the population group of the deceased. The NIDS data show that 7 305 households were interviewed and 838 households responded that at least one death had occurred within the particular household in the previous 24 months (SALDRU, 2009). Three households reported that at least one death had occurred in the previous 24 months but details about the death or deaths were reported in subsequent fields for the second decedent making it seem as though no details were recorded for the first decedent. For these households the numbers of deaths included in this analysis were the ones with reported details. On the other hand data were not weighted because it would difficult to delete deaths with inconsistencies. Appendix A1 shows the do-file that was used to merge the person data file of the heads of households to the mortality section in the household questionnaire.

There are 26 categories of relationships including cousins, great-grandparents, great-grandchildren, uncles or aunts, nephews or nieces. Other family and other non-family are trivial for this research, because the numbers of observations for these particular relationships are small or because it is difficult to assume that the age of the deceased and that of the head of the household are related. In total 948 deaths were reported by households. At most three per cent of the mortality data had the relationships between the deceased and their heads of households missing. In contrast, nearly 20 per cent of the ages at death of the deceased were

missing (Moultrie and Dorrington, 2009). This means that it is reasonable to impute using a variable with less missing observations.

Some categories of relationships that are perceived to be similar have been aggregated. In particular, absent heads of households and resident heads of households were combined and children included biological sons or daughters, step children, adopted children, fostered children and sons-in-law and daughters-in-law were also combined. Parents comprised of biological fathers or mothers, step parents, adopted parents, foster parents and the fathers-in-law and mothers-in-law. Siblings included biological brothers, biological sisters, brothers-in-law and sisters-in-law. However, relationships such as grandparents, grandchildren and husbands or wives remained specific. The proportions of deaths by the type of relationship and population group of the reported deaths are shown in Table 3.1. These proportions are stated as they were reported before considering the reasonableness of the ages at death of the deceased to the age of the head of household based on their relationship status, and therefore, Table 3.1 shows the proportions of the collated data for the reported deaths by the type of relationship and population group. The population group of the deceased was assumed to be similar to that of the head of household.

Table 3.1 NIDS reported deaths by relationship status and population group

	Africans		Coloureds		Asians/Indians		Whites		Not stated		Total	
	N	%	N	%	N	%	N	%	N	%	N	%
Heads	7	1%	3	4%	0	0%	0	0%	1	2%	11	1%
missing	15	2%	4	5%	0	0%	2	7%	4	6%	25	3%
partners	112	15%	20	24%	4	50%	5	17%	14	22%	155	16%
children	254	33%	21	25%	0	0%	11	38%	16	25%	302	32%
parents	133	17%	16	19%	3	38%	3	10%	6	10%	161	17%
siblings	86	11%	9	11%	0	0%	4	14%	10	16%	109	11%
grandchildren	77	10%	4	5%	0	0%	1	3%	6	10%	88	9%
grandparent	19	2%	1	1%	0	0%	0	0%	1	2%	21	2%
others	60	8%	7	8%	1	13%	3	10%	5	8%	76	8%
Total	763	100%	85	100%	8	100%	29	100%	63	100%	948	100%

Source: Derived from the National Income Dynamics Study (2009) data

It may be deduced that clearly identified relationships such as partners, children, parents, siblings, grandchildren and grandparents account for nearly 90 per cent of the reported data,

excluding missing relationships and 'others'. The African population accounts for 80 per cent of the reported household deaths while the Asian/Indian population reported less than 1 per cent of the deaths in the previous 24 months. There were 11 deceased heads of households in the previous 24 months before the interview, accounting for 1 per cent of the reported deaths. Unfortunately, when employing multiple imputation the ages of the heads of households are used as the independent variables, and therefore, for households where deceased heads of households with missing ages and households with other deceased people with missing ages, the missing age cannot be replaced when the age of the head of household is missing.

3.2 Age at death distribution

Ages at death of the deceased were initially plotted against ages of heads of households for each relationship status and each population group. However, data available for some of the population groups were too few to permit further analysis, and therefore non-African population groups were combined. Sex of the head of household and/or sex of the deceased were also noted to influence the ages at death distributions. For instance, female spouses are on average younger than male spouses and as a result in this analysis female headed households are separated from male headed households. In addition, children of female headed households are on average older than children of male headed households for heads of similar ages, since females have children at earlier ages compared to males, hence the average age of mothers is lower than the average age of fathers (Dorrington *et al.*, 2004). Therefore, these homogenous groups reduce noise in the data, thus reducing confounding results.

Cases where the age of the head of household was missing and/or the ages at death of the deceased were logically inconsistent with the age of the head of household were excluded in plotting scatter plots. Examples of inconsistencies are as follows; the age at death of a biological parent to the head of household that is less than the age of the head of household, or the age at death of a biological child that is more than the age of the head of household. In all these instances, the correcting hierarchy was that the age of the head of the household was assumed to be correct then the relationship status of the deceased to the head of the household, and last the age of the deceased. Ages of the heads of household were used as the independent values for imputing when using the multiple imputation method. In addition, fostered, adopted or step relationships were also interrogated to evaluate the age generational gaps to see if they were sociologically plausible. Respondents sometimes reversed the

relationships, in the case of parents and children or grandparents and grandchildren, and therefore, this was also checked and corrected where possible.

Linear regressions were fitted to the plotted ages of the deceased, plotted against the ages of heads of households. The regression coefficients for the various categories for Africans and non-Africans for which there were at least 10 deaths were tested to see if either the slopes or intercepts differed significantly from one another, if not they were amalgamated along with combinations with less than 10 deaths. Table 3.2 shows the number of cases after removing cases with incontinences and the approach employed to combine the African and non-African population groups by relationship. After removing ages with inconsistencies, missing ages at death increased to 23 per cent.

Table 3.2 Significance of intercept (*a*) and slope (*b*) of non-African and African population groups

Relationship status	Analysis Category	Population group	Number of cases	Conclusion
Partners	Male heads	African	17	Combined
		Non-African	5	
	Female heads	African	56	*a*=sig, *b*=sig
		Non-African	17	*a*=non-sig, *b*=sig
Children	Male heads	African	88	Combined
		Non-African	7	
	Female heads	African	116	Combined
		Non-African	7	
Parents	Fathers	African	22	Combined
		Non-African	9	
	Mothers	African	52	*a*=sig, *b*=sig
		Non-African	13	*a*=sig, *b*=sig
Siblings	Heads	African	74	Combined
		Non-African	8	
Grandchildren	Male Heads	African	33	Combined
		Non-African	0	
	Female Heads	African	44	Combined
		Non-African	2	
Grandparent	Grandfathers	African	3	Combined
		Non-African	0	
	Grandmothers	African	6	Combined
		Non-African	1	

'*Sig*' stands for significant. Coefficients a and b are the intercept and the slope respectively. 'Combined', describes that the African and non-African groups were combined without investigating the significance of the coefficients because of the number of cases observed in one or both of the population groups.

The regression coefficient and the intercept for the African population and the non-African population are initially tested to see if they are significantly different from zero. After testing for the significance of each coefficient in their respective linear regression models, the coefficients were then tested to see if they were significantly different from each other between the two population groups. From Table 3.2, for the population groups that were not combined, it can be seen that the coefficients for Africans and non-Africans were both significant except for one instance, the deceased partners of female headed households for non-African households. Thus, when one of the coefficients to be tested is not significantly different from zero the test for equality between African and non-African population group coefficients is not conducted, because when a coefficient is not significant in the model it means it is not important in the model and thus may be removed.

The coefficients for Africans and non-Africans that were both significant were the slopes in the case of deceased partners of female headed households and both, the intercepts and the slopes for deceased mothers. No statistical tests were conducted for population groups that are marked as combined because the numbers of cases were fewer than the required number of cases for statistical tests between coefficients to be performed. For relationships where coefficients were statistically significant, the coefficients were tested to see if the coefficients of the two models are statistically significantly different. If not, the two data sets, Africans and non-Africans, can be combined into one dataset with one regression. .An example of the detailed Stata procedures and results of the tests are in Appendix A2.

The example of the African and non-African deceased mothers is given in Appendix A2 because the intercept and the slope for both equations are significantly different from zero. It is seen that for this relationship category the African population group has 52 cases and the non-African population group has 13 cases. To test if the coefficients of the two regression models are not significantly different from one another, nested models are created, by combining the two regression models. In the process of combining the two models a dummy variable, d, and an interaction variable, w, are created. Testing whether the coefficient of the dummy variable and the interaction term are jointly zero is analogous to testing if the coefficients of the first regression are significantly different from the coefficients of the second regression, as is shown in Appendix A2.

Combining the reported groups by population group increased the number of observations for each relationship, however, the Africans dominate the composition of the categories because of the number of reported deaths for this population group. In addition, relationships for which females and males were analysed separately show different proportions by population group. Most of the relationships present strong correlations between the ages of the heads of households and the ages at death of the deceased after removing all data points that are sociologically and biologically implausible. For each relationship, except siblings, two different ages at death associations were established, based upon the sex of the head of the household and/or the sex of the deceased. The two age correlations for each relationship show similar trends for the ages at death of the deceased, increasing or decreasing as the age of the head of household increased. Figure 3.1 to Figure 3.6 show the correlation between the ages of the deceased and the ages of the heads of household by relationship after combining the Africans and the non-Africans population groups.

Figure 3.1 Correlation between grandchildren's ages at death and heads of households' ages

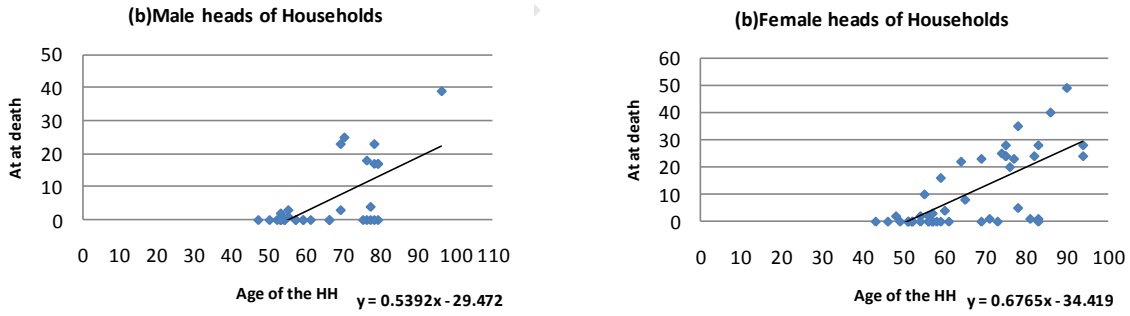


Figure 3.1 shows the ages of the heads of household and ages at death of the deceased grandchildren. The difference between the ages of the grandparent and the deceased grandchild were assumed to be at least 40 years, and cases that did not meet this condition were treated as missing or the ages inverted if the age difference was negative and it was above 40 years.

The average ages for the grandchildren are low and the age of the head of household is greater than the age of the deceased grandchild for each household as anticipated. However, there are some outliers. There are no obvious trends that can be depicted from the

relationships, which corresponds to the general assumptions about the associations between the ages at death of grandchildren and the ages of their grandfathers or grandmothers. The general assumptions are that older grandparents have older grandchildren on average, but, older grandparents also have more grandchildren of different ages. Therefore, the age at death of a grandchild cannot be independently explained by the age of the grandparent.

Figure 3.2 Correlation between children's ages at death and heads of households' ages

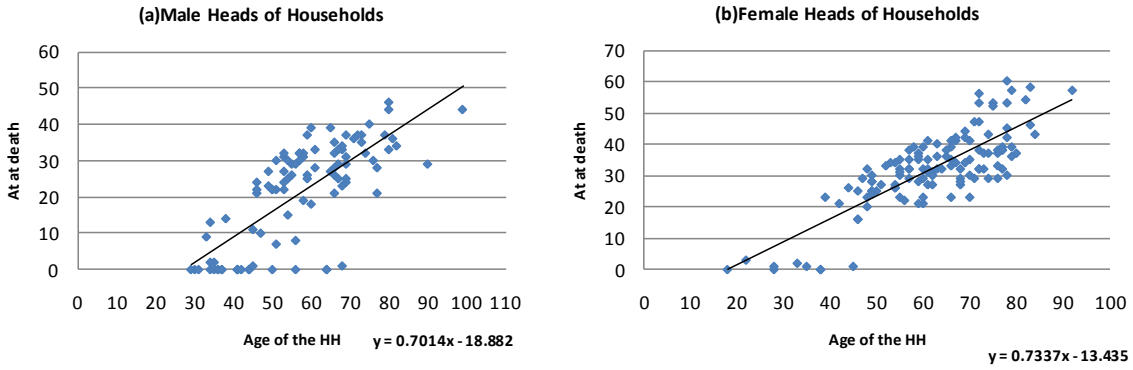


Figure 3.2 shows the relationship between ages at death of children of the head of household to that of the head of household. Female heads of households were assumed to be at least 15 years, but not more than 50 years older than their deceased children, because that is the reproductive age range. And male heads of household were assumed to be at least 20 years older than their deceased children. From this we see that there are strong positive associations between children's ages at death and those of their heads of household

Figure 3.3 Correlation between grandparents' ages at death & heads of households' ages

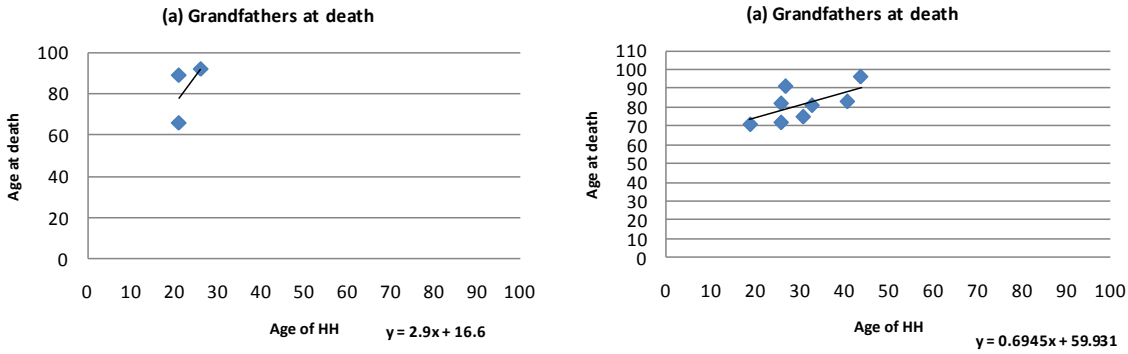


Figure 3.3 shows the correlations between ages of heads of households and ages at death for their grandparents. Cases that showed age differences that were less than 40 years have been deleted from the analysis as 40 years is the minimum expected generational gap. Ages at death of grandfathers increase with increasing ages of heads of households, but the data points are too few to seriously consider this relationship. On the other hand ages at death of grandmothers show a similar trend and have more data points to consider.

Figure 3.4 Correlation between partners' ages at death and heads of households' ages

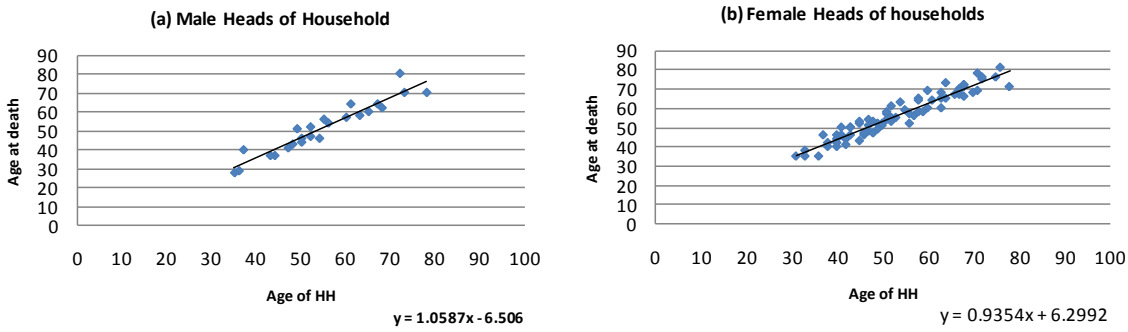


Figure 3.4 shows the relationship between the ages of the partners and the head of households. Couples were assumed to have an age difference of not more than 10 years, and therefore cases that had an age difference of more than 10 years were treated as missing. From this we see that in general ages of partners at death increase as the ages of the heads of households increase. Thus, there is a strong linear relationship between the ages at death of the deceased partner and the ages of the heads of households.

Figure 3.5 shows the correlation between the ages at death of parents and the ages of their heads of households. Similar age differences to those of the ages of deceased children and their heads of households were excluded. That is the age difference between the deceased father and the head of household was assumed to be at least 20 years and those ages that reflected an age difference that is more than 20 but negative, their ages were reversed. Only two households had this situation. The two age differences were of a plausible magnitude but were negative, suggesting that these relationships may have been inverted, reporting the relationship of the head of household to the deceased rather than the other way round. The households had reversed relationships because for these particular households the relationship codes taken from the mortality section of the household questionnaire and relationship codes shown in the household roster were different. On the other hand deceased mothers are assumed to have an age difference that is at least 15 years but not more than 50 years between them and the head of household.

The ages at death of parents increase as the age of heads of households increase. Fathers are expected to be older than mothers on average, but this is not obvious from the scatter plots because data are plotted separately for male and female heads of household. When heads of households are of similar ages, the average age difference is significant because the mean age at marriage of women is lower than the mean age at marriage for men (Dorrington *et al.*, 2004). Deceased parents of heads of household of similar ages may be assumed to have deceased fathers older than deceased mothers because it is assumed that they were born within marriage, because Foote *et al.* (1993) identify marriage as one of the proximate determinants to fertility.

Figure 3.5 Correlation between parents' ages at death and heads of households' ages

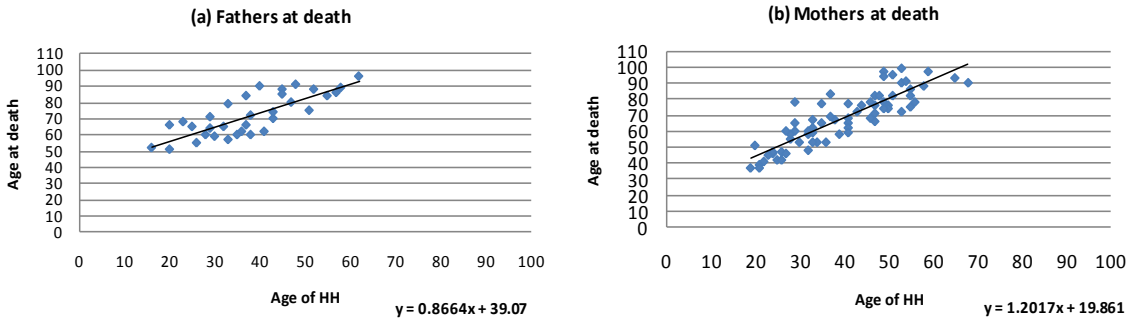


Figure 3.6 Correlation between siblings' ages at death and heads of households' ages

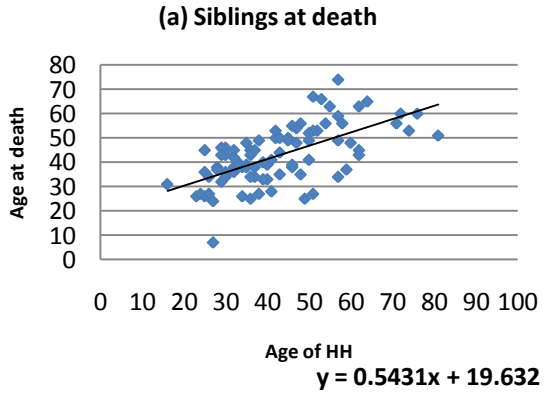


Figure 3.6 shows the association between the ages at death of the siblings of the head of household and the ages of the heads of households. However, when siblings are born of the same mother, it is assumed that their age difference should not be more than 35 years because of the limited fertility period of women. Therefore, age differences between the deceased sibling and the age of the head of the household that were more than 35 years were deleted and the cases treated as missing. Siblings may be older or younger than the head of household and the gender of the sibling does not determine the age of the sibling in relation to the head of household. Thus Figure 3.6 shows points that are scattered and no obvious pattern is depicted.

3.2.1 Missing mechanism

In section 2.1, the missing mechanism for the ages at death was discussed, that is, whether the missing ages are completely missing at random, missing at random or not missing at random.

This is tested by inspecting a table showing the percentage of missing ages by the relationship of the deceased to that of the head of household as shown in Table 3.3. This table also shows the total number of cases in each relationship category when cases without population group stated were added and also after deleting all inconsistent cases. Percentages of missing ages at death by relationship status which are not similar imply that the relationship of the deceased to the head of household influences the missing ages, and therefore that the ages at death are not completely missing at random. Primarily, an investigation of this type can be made by using a continuous variable such as the age of the head of household and the percentage of missing ages at death can be plotted against the ages of the heads of households. If the percentages are similar then a case deletion, whereby observations are removed from the analysis by deleting cases that have been not completely observed for all variables, would be suitable. The data show that the missing ages at death differ by relationship status and age of the head of household, and therefore that it can be concluded that missing ages at death are not completely missing at random. Thus a case deletion would not yield legitimate or valid results.

Table 3.3 Percentage of missing ages at death by relationship status

Relationship status	Analysis Category	Complete cases	Total	Per cent missing
Partners	Male heads	24	42	42.9
	Female heads	74	111	33.3
Children	Male heads	126	175	28.0
	Female heads	98	117	16.2
Parents	Fathers	33	55	40.0
	Mothers	68	103	34.0
Siblings	Heads	86	106	18.9
Grandchildren	Male Heads	33	35	5.7
	Female Heads	47	51	7.8
Grandparent	Grandfathers	8	6	50.0
	Grandmothers	3	14	42.9

3.3 Multiple Imputation

This research employs the STATA programs for Multivariate Imputation by Chained Equations (MICE) described in Royston (2004) to undertake multiple imputations as proposed by Rubin (1987), the sub-routines were downloaded from Internet Documents in Economics Access Service (IDEAS)[1]. In particular, *uvis* which performs univariate imputation sampling for missing values of the *y*-variable was used, together with the regress command to implement the linear regression. The algorithm that is implemented by the *uvis* command first regresses the non-missing *y*-values on complete *x*-values. As a result values for the coefficients and residual variance are established. Second, a random value σ^* (sigma star) is drawn from the posterior distribution of the standard deviation of residuals. Third, a value of β^* (beta star) is drawn at random from a posterior distribution of beta conditional on σ^* thus allowing for uncertainty in beta. Fourth, the values of β^* are then used to predict values for the missing values of the *y*-variable. Thus, the values imputed are derived from β^*, σ^* and the covariates. Imputation using linear regression assumes that the *y*-variable is normally distributed given the covariates, however, when linear regressions are not being employed to replace missing observations, another kind of distribution can be observed from the dataset and used to impute (Royston, 2004).

Ages at death have a skewed distribution, therefore, the default version of *uvis* cannot be employed and hence the *match* option for *uvis* had to be employed. The *match* option is suitable for continuous data that are not normally distributed and produces better imputations than the default method in all situations that data are continuous and not normally distributed. As a result of applying the *match* option, in the fourth step in the preceding paragraph is then changed, so that values are imputed through predictive matching using the matching criterion that draws imputations randomly from a posterior distribution of the missing values of the *y*-variable, conditional on the observed values and the values of the *x*-variable (Royston, 2005).

For this research, three sets of imputed values were created, using the minimum criterion suggested by Rubin (1987) for data with less than 20 per cent missing observations. The three complete datasets created can be used for point estimators of the mean and the variance for the complete dataset. These point estimators are calculated as the averages of the means and variances of the datasets created. When one complete dataset is desired, the mean

[1] The URL for the sub-routines on IDEAS website is *http://ideas.repec.org/c/boc/bocode/s446602.html*

for each case of the imputed values could be used. However, using the mean reduces the variance, therefore, one complete dataset was created by randomly selecting an imputed value for each case from the three created datasets. See Appendix A3 and Appendix A4 for the Stata do files that were written to run multiple imputations and run to select one dataset to represent the three datasets respectively. The first do-file generates the data by formulating a regression equation on the observed complete cases and uses this regression equation to impute the missing dependent values using the observed value for each case.

3.4 Comparisons between Statistics South Africa's hotdeck and SOLAS hotdeck

Hotdeck imputation is used by way of comparison. The employment of the SOLAS method was decided upon after attempting unsuccessfully to create a Stata do-file to reproduce Statistics South Africa's hotdeck, because it was created using CSpro software. The results of employing SOLAS hotdeck imputation software were compared to hotdeck imputations done on the Census 2001 age at death data by Statistics South Africa to see if the SOLAS hotdeck could be used as a proxy for Statistics South Africa's hotdeck method.

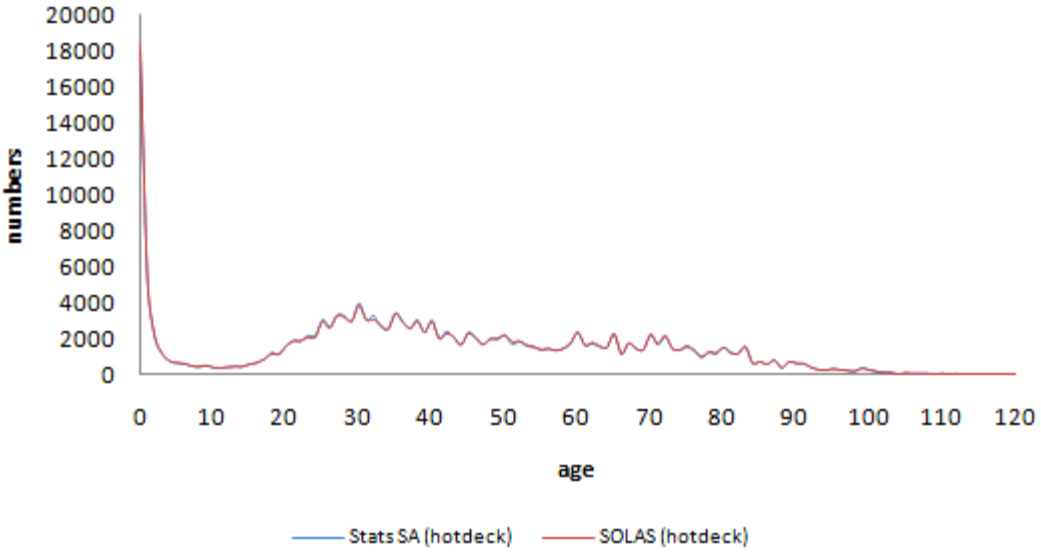
SOLAS software for missing data analysis[2] employed as a proxy to Statistics South Africa's hotdeck is written by Statistical Solutions, which is a software development company, in conjunction with Prof Donald B. Rubin, to provide researchers using data with missing values various imputation techniques.

Imputations were done separately, for those deceased that had missing responses on the pregnancy status of the pregnancy related questions and those deceased that had valid responses on the pregnancy related mortality question. A missing response on the pregnancy status at death means that the deceased is either a male or the deceased was a female aged younger than 12 years or older than 50 years. On the other hand, a valid response, which could be 'Yes' or 'No', limited the ages of women to be between 12 years and 50 years. The pregnancy status of the deceased at death was then used as a sorting variable in addition to the population group variable to implement the hotdeck in the same way it was used for imputation by Statistics South Africa.

[2] SOLAS software is available at *http://www.statistical-solutions-software.com/products-page/solas-for-missing-data-analysis/*

To show the adequacy of SOLAS hotdeck as a proxy to Statistics South Africa's hotdeck, the age distribution after SOLAS hotdeck is compared to the age distribution after Statistics South Africa's hotdeck. However, since the missing age at death data imputed by means of hotdecking are nearly eight per cent of the complete dataset, the effect on the distribution of the ages at death data may be expected to be minimal. Figure 3.7 shows the distributions of the two datasets on the same axis.

Figure 3.7 Age at death distributions after SOLAS and Statistics South Africa's hotdeck



The absolute difference and relative fractions were calculated to show the difference around zero and one respectively. The absolute differences are erratic revealing high differences for age zero and for ages between 20 years to 55 years. On the other hand, the ratios are all within 10 per cent. Appendix A5 shows the absolute differences and the ratios of the two datasets on the same axis. The high absolute differences may be a result of many observations in these age groups with similar covariate values. Thus, most of the imputed ages are imputed from ages that are recurring more than the others.

3.5 Hotdeck in Stata

The subroutine for hot deck imputation as proposed by Mander and Clayton (2007) first tabulates the missing patterns by rows and columns, which are the records and the sets of

variables, respectively. For records, the set of variables with non-missing values are then employed as covariates for the variables with missing values. The covariates are variables that are used to match values in their fields with values in complete lines so that the missing value or values may be imputed. When a record has at least a single value missing, the STATA hot deck subroutine recognises the record as a line with a missing value or values and complete records are recognised as complete lines. Thus the lines are categorized as *nmiss* and *nobs* for an incomplete line and a complete line, respectively. The imputation method is executed stochastically or randomly, and, the hotdeck is run several times within a multiple imputation sequence. When the hotdeck is executed stochastically, it means that records with complete data are sampled with replacement and used to replace missing values in records with missing data by matching the covariates. The multiple imputation sequence means that several variables are simultaneously imputed. The number of times that the hotdeck is run depends on the percentage of missing information and the type of data, but Mander and Clayton (2007) suggest that five times ought to be sufficient when the dataset does not have many variables with missing data. When the dataset has many variables with missing data, it is likely that many of the rows of data will contain at least one missing value.

The major options used are; *'by'*, which creates strata that demarcate the categories for which missing lines are going to be replaced by complete lines, in particular the population groups are the strata and *'using'*, which creates the imputed datasets. Missing data are perceived to be missing completely at random (MCAR) or missing at random (MAR) within the strata created, otherwise when there is a pattern within the strata the results are distorted. Data were composed of reported ages at death, month of death, year of death, sex of the deceased and whether the cause of death was natural or accidental/violent. Some of these fields were reported as *'unknown'* or they were reported as *'respondent refused to answer'*. In these circumstances the values in these fields are recognized by STATA as *'not missing'*, therefore the fields with these values had to be replaced by a *'missing'*, in order to be able to employ the hotdeck successfully. Unfortunately there is no variable which reports whether the deceased woman was pregnant or not, and hence the imputations of the ages at death are similarly treated for all females.

3.6 Apportionment of missing ages at death

This method is usually employed by researchers to replace missing observations in the absence of leading information, such as other variables to assist in imputing. There are two methods of apportioning the ages. Missing ages are usually distributed across the whole age range or distributed for ages above 20 years because respondents are assumed to know children's ages better than ages of adults. In this research the latter method is used on *a priori* bases, however, the former method is also investigated.

3.7 Vital registration data

Mortality data from registered death notifications for the year 2006 in five year age groups are used in this research as the benchmark against which to evaluate the quality of imputations assuming that the completeness of registration of deaths is the same for all age groups. The dataset for 2006 was used because it was the year that most closely matched the data from the survey. The proportions by age group are calculated and these proportions are compared to the proportions that are calculated for the datasets which included imputed ages at death created after employing the following methods; multiple imputation method, Statistics South Africa's hotdeck, STATA hotdeck and simple apportionment.

Differences in completeness of death registration by age are important because they affect the frequency distribution of ages at death, hence the proportional distribution is also affected. However, for this research, death data are assumed to be uniformly under-reported to enable proportion comparisons between the vital registration ages at death data and the NIDS complete data with imputed ages at death. This assumption can be made because, Darikwa (2009) observes that, death reporting for children has improved and completeness of reporting has risen close to a point where it is comparable to the completeness of reporting of adult deaths.

Chapter 4 Analysis and Results

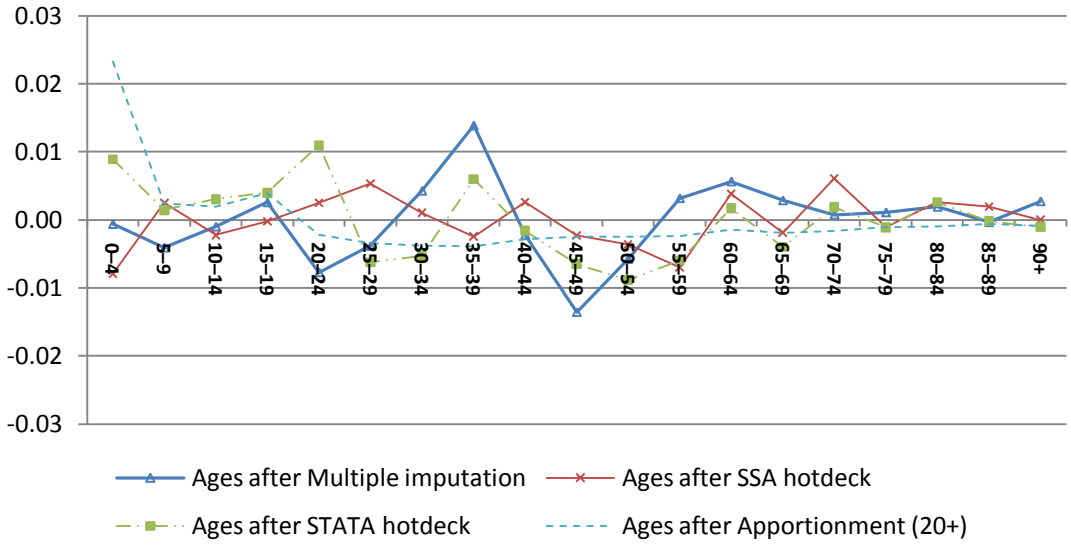
The results of the methods applied in Chapter 3 are presented in this chapter. Initially the results for each method are discussed in comparison with the observed ages at death for deaths for which plausible ages at death were recorded, after that the vital registration reported deaths are used as the standard, assuming that under reporting is constant with age.

4.1 Comparisons between imputed datasets and plausible observed data

The quality of imputations was first determined by the similarities between data with reported ages, that is, the observed data, and the complete dataset which includes the imputed values. The observed data referred to here and hereafter is the corrected data after removing implausible ages and the original dataset is referred to as the observed dataset before editing. For multiple imputation scatter plots were constructed to present a graphical view of the relationship between the imputed data and the ages of their heads of households. The Kolmogorov-Smirnov test was conducted to evaluate if the two distributions were significantly different from each other. In addition, empirical density estimates were constructed to enable illustrative evaluations of the Kolmogorov-Smirnov test.

The proportional differences between observed data and each of the imputed datasets using multiple imputation, Statistical South Africa's hotdeck, simple apportionment and the Stata hotdeck were plotted. The proportional differences shown in Figure 4.1 are the differences between the proportions by age group in the observed dataset and the proportion in each dataset created. Most of the proportional differences fluctuate between -1 per cent and 1 per cent, save for proportional differences after multiple imputations and after simple apportionment conducted to the ages greater than 20 years. Multiple imputation exhibits glaring differences in the middle age groups, unlike simple apportionment that has the proportional difference greater than one in the 0-4 age group. Proportional age differences calculated for simple apportionment that assigns ages to the whole age range will have all differences equal to zero, meaning that the distribution of the observed dataset and the complete dataset will be the same. This comparison is thus not included in Figure 4.1.

Figure 4.1 Differences in proportions by age – proportions after imputation less proportions of those with recorded ages



4.1.1 Multiple Imputations (MI)

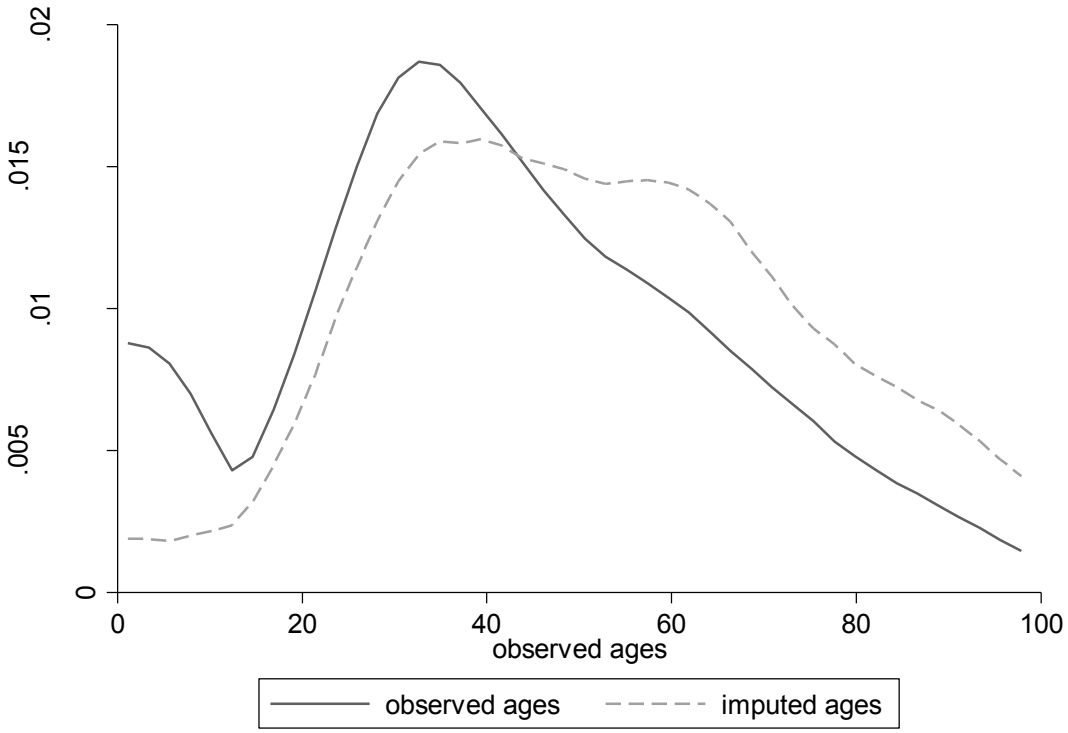
Missing ages of the deceased were imputed in accordance to their relationship category, thus the results were initially analysed in the relationship category of the deceased to their head of household. Appendix A6 show the bivariate scatter plots of the complete ages at death data against the age of the head of household by relationship category. The scatter plots show that there are no marked differences between the imputed and the observed ages at death. The scatter plots show that the imputation procedure based upon the relationship of the deceased to head of the household results in age imputations that are not radically different from the reported or observed ages in the relationship category. The results are in contrast to the results expected from the initial scatter plots between the observed ages of the deceased and those of heads of households in their relationship categories, which show weak linear relationships. The results anticipated were that the imputed ages will be distributed differently from the reported ages at death because the imputations were based on weak linear relationships.

The relationship categories were then combined so that the observed dataset before editing, the observed dataset after deleting implausible relationships and the complete dataset after imputation could be compared. The frequency distributions of the age groups for the three datasets were compared to see if the editing to remove implausible ages was necessary,

because this could be hidden when densities are used. Fluctuations in the frequencies of the age groups after imputation do not exactly depict distribution patterns observed in the other two datasets. However, it cannot be said that the distributions are markedly different as several age groups at death show similar patterns. The proportional differences in Figure 4.1 show evident differences in the 35-39 year age group and the 40-44 year age group between the corrected observed data and the data imputed by employing multiple imputation. This is because the concentrations of missing ages are within relationships that are inclined to impute ages in these age groups, such as deceased partners of heads of households. See Appendix A6 for the scatter plots showing observed data after removing implausible data points and imputed data. It can also be observed in Appendix A6 that most of the imputed ages were imputed for relationships inclined to the middle and older ages, for instance deceased partners.

The Kolmogorov-Smirnov test was used to check if the null hypothesis, H_0, that the distribution of the imputed values and the observed values were the same, was true. At the five per cent level of significance the test concludes that there is a significant difference between the distribution of the observed ages and the distribution of the imputed ages (See Appendix. A7). With an exact p-value less than 0.001, it means that the distributions are statistically significantly different, which is further supported by the empirical density plot in Figure 4.2. The empirical density plots are pictorial representations of the Kolmogorov-Sirmnov tests, by approximating the density $f(x)$ from observations on x. In addition, under the missing at random assumption, data could be missing, dependent on another observed variable, therefore, the distribution of the reported values or observed values may not be expected to be the same as the distribution of imputed values, however, a marked difference calls into question the suitability of the imputation procedure employed. For instance, questions on the suitability of the imputation procedure may arise when the reported or observed ages are unimodal and the imputed ages are bimodal. Figure 4.2 shows the distributions of imputed ages and the observed ages for the combined relationships or the empirical density estimates after multiple imputation. There is a difference between the density distributions of the imputed ages and the corrected observed ages after removing implausible ages. The corrected observed ages have two humps, one in the lower ages and the second one in the higher ages, however, this is not the case with the imputed values that show one hump in the middle ages and a kinky shape between the ages of 60 years to 80 years. It can be said that imputation was inclined to the older ages as described above.

Figure 4.2 Empirical density estimates: deaths with age reported and deaths after multiple imputation



The empirical density estimates are also known as kernel density estimates. The kernel is the function that determines the weights that are between 0 and 1 with respect to the distance from the center of the bar or window that is identical to the bars of a histogram. The Epanechnikov kernel was employed because it minimises the integrated square error more efficiently in comparison to other kernels such as the Bweight, Cosine and Gausian (StataCorp., 2009). The width of the bar or window was optimally chosen by the Stata software. Figure 4.1 shows that the shape and the level of the density curves for observed and imputed ages are slightly different. When the densities of the observed and the imputed values are similar, under the assumption of missing at random, the imputation procedure is regarded as suitable. However, multiple imputation that incorporates the relationship of the deceased can be seen to have transformed the distribution of the dataset. The increase in missing observations increases the rate of missing data as shown in Table 2.1, which shows that the

efficiency of estimates may only increase when more imputed datasets are created, when the rate of missing information is high.

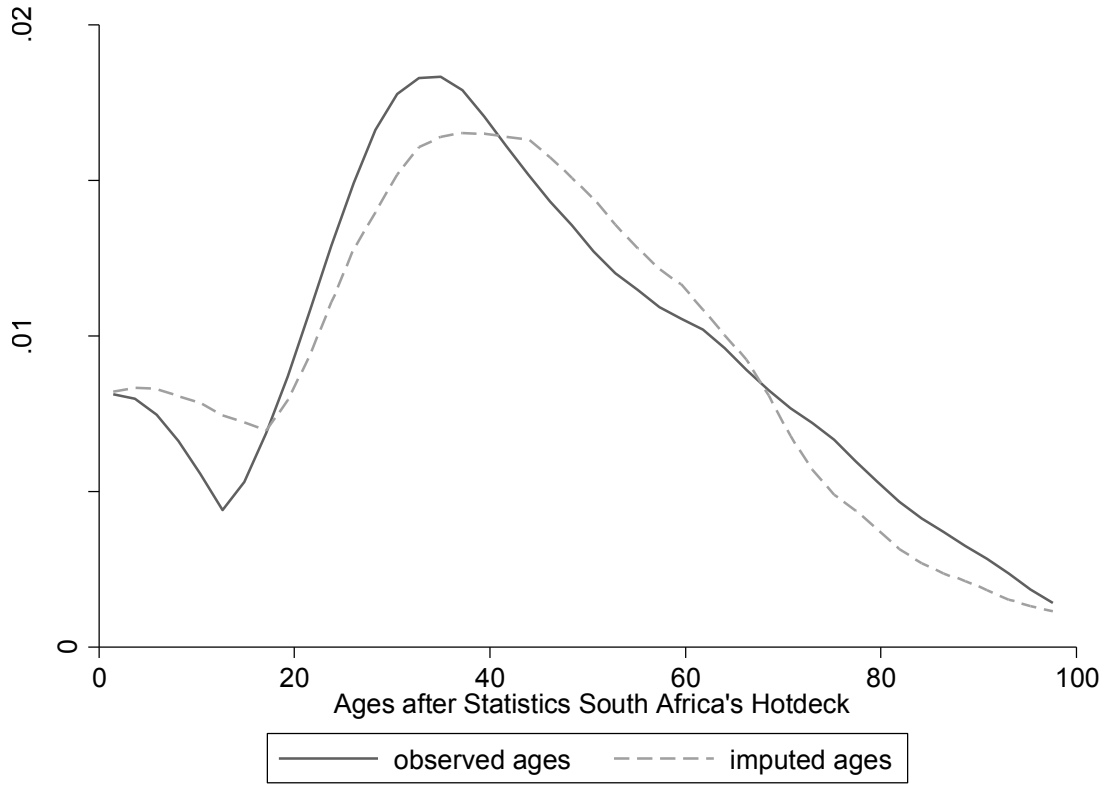
4.1.2 Statistics South Africa's hotdeck

The SOLAS hotdeck routine was employed as a proxy for Statistical South Africa's hotdeck. This was done after comparing the two distributions using the 2001 South African Census age at death data and concluding that the two distributions were sufficiently similar. The detailed procedures and conclusions are presented in section 3.5.

Frequency distributions of observed ages at death data and complete ages at death after imputation were compared. They show different distributions in the middle ages, with the peaks occurring in different age groups. In contrast, the age groups above 60 years show no significant differences perhaps because of the small numbers of observed ages at death in these age groups. On the other hand, the proportional differences in Figure 4.1 fluctuate between -1 percent and 1 per cent making the imputation procedure the one with least erratic proportional differences. Age groups 45-49 to 55-59 have negative proportional differences, which show that the distribution after imputations is peaked for these ages.

Figure 4.3 shows the empirical densities of the imputed ages at death and the observed ages at death. The curves are similar, which confirm that the imputed ages at death and the observed ages at death have similar distributions, save for overestimating the densities for the ages below 20 years and underestimating the densities for ages up to 45 years, then slightly overestimating up to 65 years. The density estimate differences observed are insignificant, with the Kolmogorov-Smirnov test giving an exact p -value of 0.6. It can thus be concluded that the imputed ages and the observed ages have similar distributions (See Appendix A7).

Figure 4.3 Empirical density estimates: deaths with age reported and deaths after Statistics South Africa hotdeck

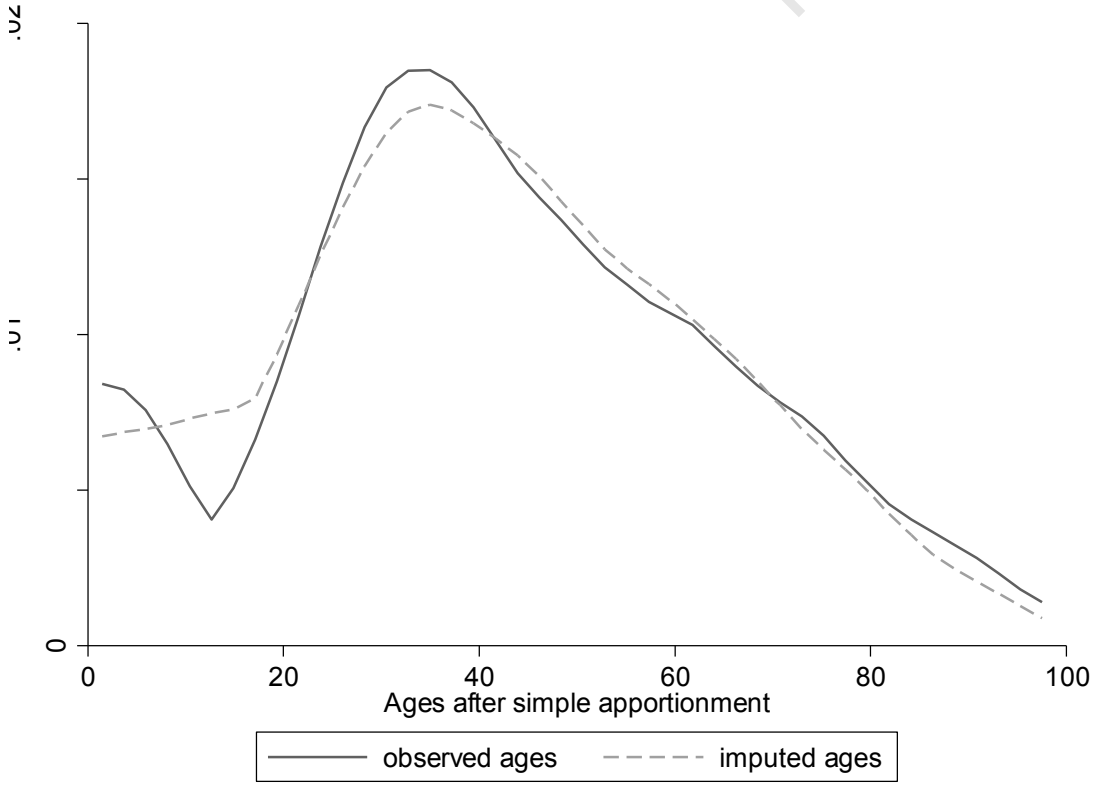


4.1.3 Apportionment Method

One of the limitation of the method is that the results for most of the subgroups are often irrational numbers (Saari, 1978). That is, the results need to be rounded off to an integer.. When a subpopulation, n_i has more observations when compared to other groups, the apportionment method allocates more observations to this group from the group with missing values. The assumption made is that more of the missing values should have been in the group with more observations. This assumption does not hold when data are assumed to be missing completely at random or missing at random because by applying the method a ratio of the observed data for the particular group to the total is used to distribute the number of missing observations. The distribution of a number of deaths by age for the observed ages at death data and the complete ages at death data after simple apportionment that distributes the ages across all ages reveal that the complete dataset has a distribution similar to the observed dataset except that it is shifted vertically depending on the observed frequency of each age which is similar to the findings of Saari (1978).

The empirical density estimates in Figure 4.4 show that the densities of the imputed ages after simple apportionment and the observed ages are similar. The difference in the shape of the curves in the ages below 20 years is the result of the kernel that smoothes the curves and the effect of rounding off the decimals to obtained number of ages allocated to a particular age . The similarity between the curves is also confirmed by the Kolmogorov-Smirnov test, at the 5 per cent level of significance concludes that there is not enough evidence to reject H_0 , and therefore, with an exact p-value equal to one, and we conclude that the datasets have the same distribution (see Appendix A7).

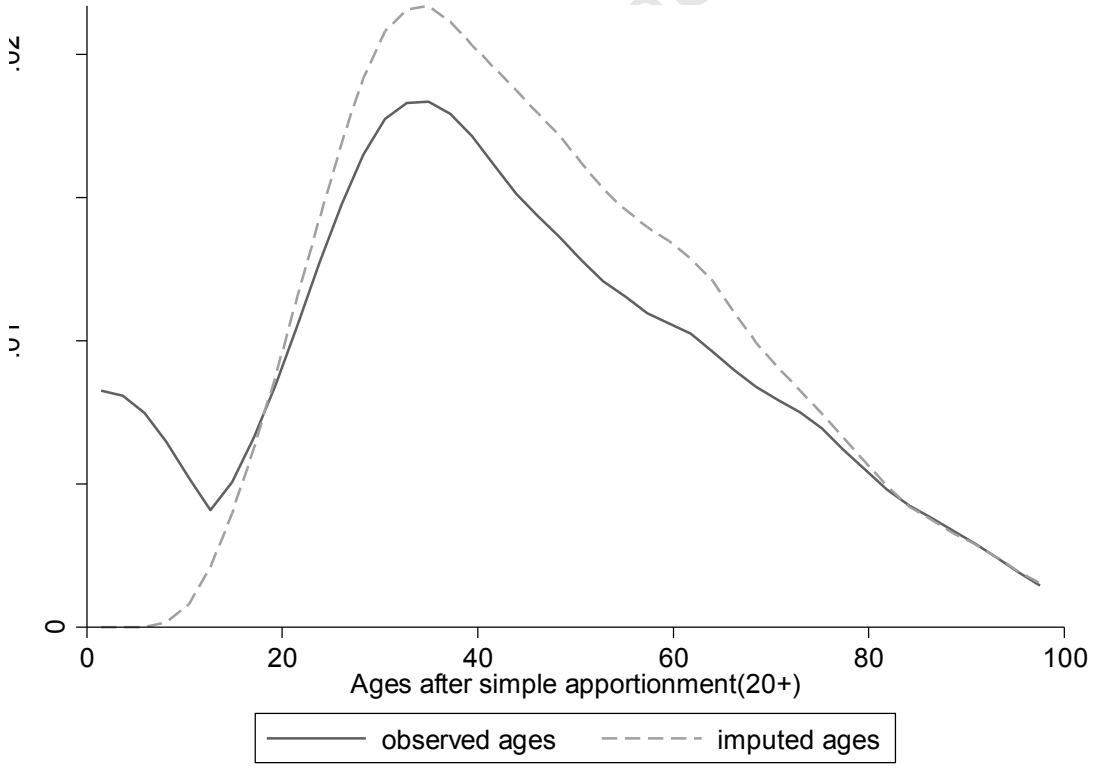
Figure 4.4 Empirical density estimates: deaths with age reported and deaths after simple apportionment



Missing ages at death may also be distributed to adult ages only, that is 20 years and above. By allocating missing ages to the adult ages, the distribution of the complete ages and the observed ages differ significantly. The frequency distribution for the younger ages below 20 years is unaltered, however the proportions are affected because age groups above this age have more ages allocated to them.

The empirical density estimates in Figure 4.5 show that there are glaring differences between the observed age distribution and the complete age distribution with apportioned ages. The distribution for the complete dataset is peaked in the middle ages and for ages below 20 years the empirical densities are lower than the empirical densities of the observed dataset. The Kolmogorov-Smirnov test shows a similar result, with a p-value of 0.001, and thus at the 5 per cent significance level, H_0 is rejected and it is concluded that the distributions of the observed dataset and the complete dataset are different (see Appendix A7). In comparison with other empirical density plots analysed, the empirical density estimates for simple apportionment to adult ages presents the most evident differences from the empirical density estimates of the observed dataset.

Figure 4.5 Empirical density estimates: deaths with age reported and deaths after simple apportionment (20 years and above)

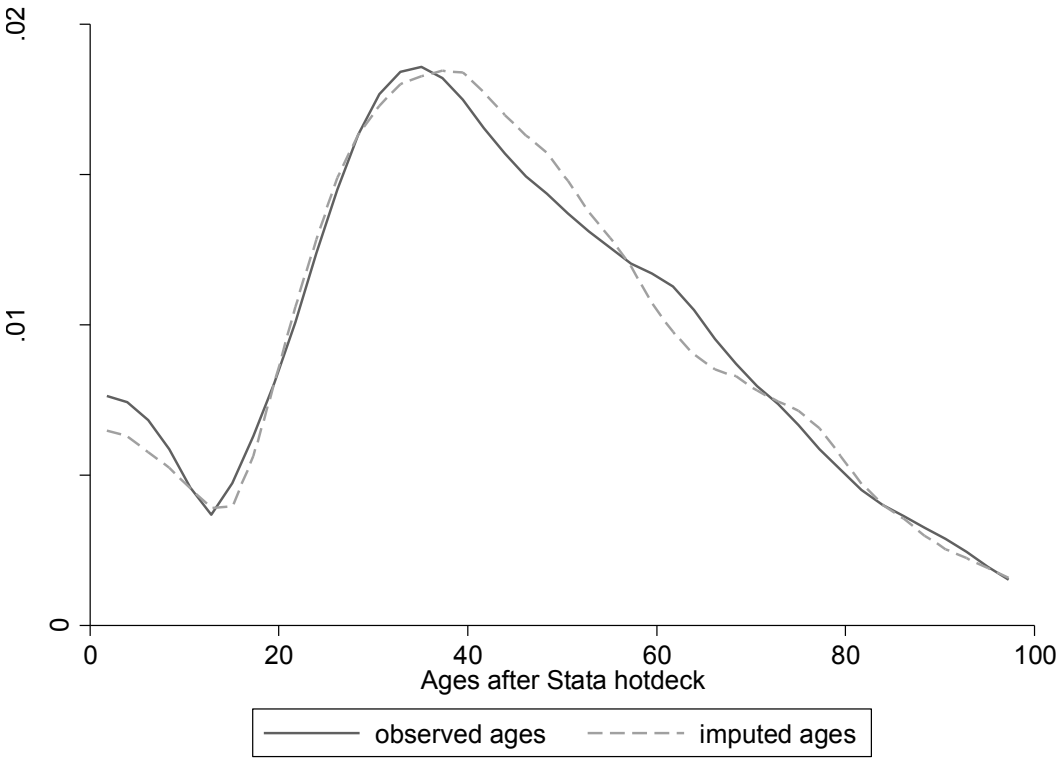


4.1.4 Stata hotdeck

Implementing the Stata hotdeck to replace missing ages using a subroutine written by Mander and Clayton (2007) results in a frequency distribution that does not resemble the pattern depicted in the observed data. The frequency distribution of the ages at death after imputation is more peaked in the middle ages from age 25 years to age 50 years. The proportional differences in Figure 4.1 show positive differences for the 25-29 age group and negative differences up to the 55-59 age group, save for the 35-39 age group. The observations on the proportional age differences are in consensus with the observations on the frequency distribution that show a peaked distribution for the complete imputed dataset. In Appendix A8, the Stata hotdeck code that was used is given. Five iterations were done to increase the efficiency of the estimates.

On the other hand, the empirical density estimates in Figure 4.6 show similar densities for imputed ages and observed ages with small differences for the ages below 20 years and the ages above 40 years. The Kolmogorov-Smirnov statistic also confirms that the two age distributions are similar. The exact p-value is 0.8, and therefore, there is no significant evidence to reject H_0 at the 5 per cent level of significance and it is concluded that the imputed ages and the observed ages have similar distributions (See Appendix A7).

Figure 4.6 Empirical density estimates: deaths with age reported and deaths after Stata hotdeck



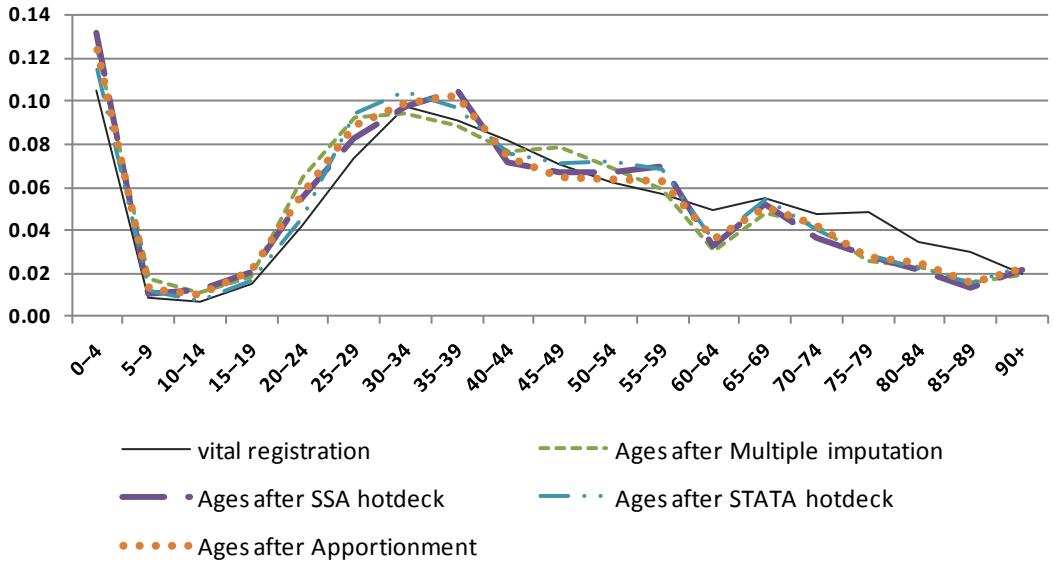
4.2 Comparisons between complete datasets with imputed ages and the vital registration

Proportions are calculated for the complete age at death dataset created using the multiple imputation method, Statistics South Africa's hotdeck, Stata hotdeck and simple apportionment. The ages are categorized into five year age groups to match the presentation of the 2006 registered deaths by Statistics South Africa. The subsequent section presents comparisons between the proportions calculated using the vital registration and the proportions calculated using each of the methods discussed in turn.

Figure 4.7 shows the proportions calculated after employing the methods stated above and from the vital registration data. When the proportions are plotted on the same axis to enable comparisons between all the methods of imputations, it can be observed that the proportions for the age groups above the 55-59 for all the methods and the observed data are below the proportions for the vital registration data. Therefore, it may be concluded that the shortfall in proportions originates from the observed data, which is the National Income and

Dynamics (NIDS) dataset. Significant differences are observed for the age groups below the 55-59 age group because the proportions for different methods are fluctuating around those observed for the vital registration. However, there are two curves that have the same distribution, because they coincide, since they have equal proportions calculated by age group. These are the proportions calculated from the observed NIDS data and the proportions calculated after using simple apportionment that assign ages to the complete age range.

Figure 4.7 Proportions of deaths by age: vital registration compared with estimates after imputation



4.2.1 Multiple Imputation (MI)

The proportions in comparison with those of the vital registration data are higher for the age groups below the 30-34 age group on average and lower for the age groups above this age, save for 45-49 age group and 50-54 age group as shown in Figure 4.7. In general, the proportions reveal that there are more deaths with ages below the 30-34 age group in comparison to the ages above this age group. This is a result of imputing ages at death based on the relationship of the deceased to the head of household. Thus, when more deaths are missing in relationships with the age at death of the deceased that is less than the age of the head of the household then this characteristic is depicted in the imputed ages. For instance, the NIDS age at death data by relationship had more ages at death missing for children and

grandchildren compared to other types of relationships such as grandparents or parents that allow ages at death that are replaced to be above the ages of the household heads. Therefore, the relationship variable needs to be scrutinised to see if it is representative of the population.

The proportions in Figure 4.7 are slightly lower in comparison with those from the vital registration data from the 30-34 age group to the 40-44 age group. Multiple imputation produces a distribution that is different to the that of the other methods which have a peaked distribution for these age groups. However, the underestimation of proportions in the older ages observed for all methods investigated means that the discrepancy is not within an imputation procedure but the observed data.

4.2.2 Statistics South Africa's hotdeck

The proportions in Figure 4.7 by age group for the ages at death after employing Statistics South Africa's hotdeck are slightly above the vital registration proportions up to age group 35-39 and from this age group, proportions are erratic up to the 55-59 age group. After the 55-59 age group the proportions calculated after employing Statistics South Africa's hotdeck are lower in comparison with the vital registration proportions. This shows that a better fraction of the ages at death are below the 35-39 age group for the imputed dataset as compared to the age at death of vital registration data. Statistics South Africa's hotdeck has got the most peaked distribution from the 30-34 age group to the 40-44 age group. The observation confirms the previous investigation by Dorrington *et al.* (2004). In comparison to other methods investigated, the deviations from the vital registration are similar except for multiple imputation in the middle ages which is lower than the vital registration. Comparing Statistics South Africa's hotdeck to the methods investigated in this research shows that the proportional distributions are similar, save that the proportional distributions after multiple imputation are lower than those of the vital registration.

4.2.3 Apportionment Method

The proportions presented in Figure 4.7 are for simple apportionment, which allocates missing ages across the whole age range, because simple apportionment that allocates ages to adult ages had empirical density estimates that were significantly different to the observed data. Therefore, the method was not investigated for external inconsistencies because the internal inconsistencies were large. The distribution of the proportions is similar to that of the

Statistics South Africa's hotdeck across the whole age range and clearly different to multiple imputation in the middle ages. The proportions in comparison to the vital registration proportions are significantly higher in the 0-4 age group and from the 15-19 age group to the 30-34 age group and lower thereafter.

4.2.4 Hotdeck in Stata

The proportions calculated Figure 4.7 after imputing missing ages at death by employing a Stata hotdeck sub-routine programmed by Mander and Clayton (2007) are slightly higher than the vital registration proportions up to the 20-24 age group, (Zang and Walker, 2008)then significantly higher from the 25-30 age group to the 35-39 age group, thereafter the proportions are erratic, showing lower proportions from the 60-64 age group. The lower proportions for the ages above 60 years are inherent in the observed data. The high peak of the proportions in the middle ages was also observed for other methods investigated in this research, except for multiple imputation.

4.3 Conclusion

Internal inconsistencies for each method were evaluated by comparing the distributions of the imputed ages and observed ages, the results show that all the distributions of the imputed values and observed values were not significantly different from each other, save for the multiple imputation method. In comparison to the proportion of deaths by age from the vital registration all the methods deviated in a similar manner, especially for the older ages, above 55 years, where all the distributions have proportions lower than those of the vital registration. The reason for a similar deviation for all methods is not clear but it may be something inherent to the NIDS dataset or the vital registration dataset.

Chapter 5 Discussions and Conclusions

The purpose of this research was to compare the methods which are extensively employed for imputing age at death, namely, Statistics South Africa's hotdeck, Stata hotdeck and simple apportionment methods, to a technique that incorporates the relationship of the deceased to the head of household variable. This variable was included in the National Income and Dynamics Study (NIDS) data. The comparison was designed to see if the inclusion of the relationship variable of the deceased to the head of household would assist in allocating ages to those deceased with missing ages at death.

The NIDS dataset had 18.5 per cent missing ages at death while the missing relationships were less than 3 per cent. Thus in the case of a missing relationship of the deceased to the head of household, a missing age could not be replaced using multiple imputation method that incorporates relationships. Hence, nearly 2 per cent of the dataset was lost. Furthermore, through assessing the reasonableness of the age at death of the deceased and the age of the head of household based on the their relationship status, some ages at death for the deceased were treated as missing, leaving the age of the head of household to be used as the independent value to impute the missing age of the deceased under multiple imputation. Therefore, eventually 23 per cent of the ages at death data were regarded as having missing ages at death totalling to 96 missing ages at death that needed imputation.

The scatter plots employed to show the correlation between the age of the deceased and the age of the head of household did not initially show biologically and socially plausible age differences by relationship. This revealed that respondents misreported either the age of the head of household or the age and the relationship of the deceased to the head of household. After data cleaning, the remaining points showed plausible age relationships, but relationships such as deceased grandparents were left with only 3 complete cases for analysis. This reduced the power of the analysis. Therefore, the dataset used had a skewed distribution of deaths by relationship. This limited the possibility of exploring the usefulness of these relationships.

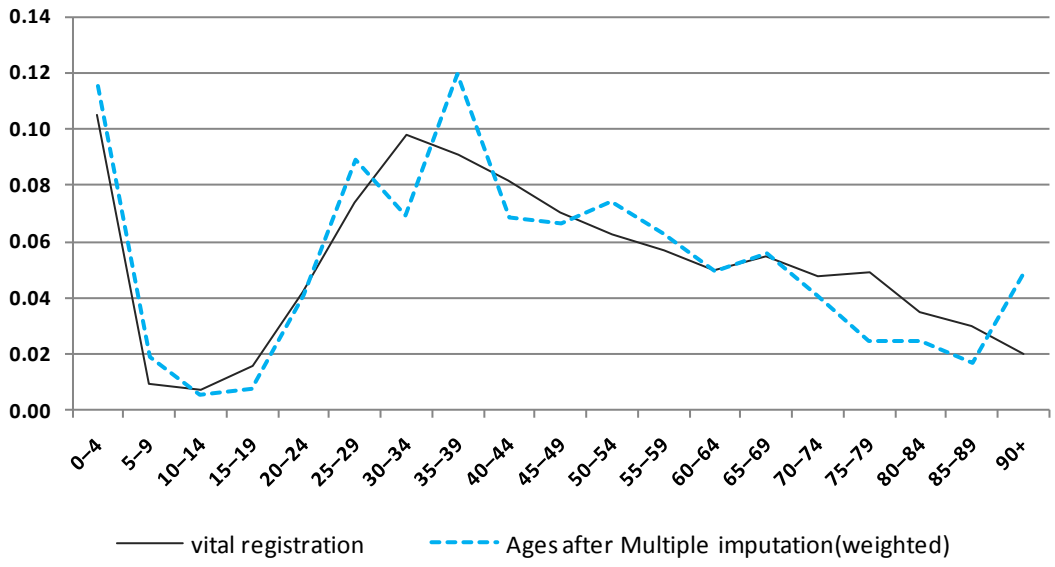
The questions asked of the main respondent of the household about mortality in the previous 24 months in the NIDS survey offered the right variables for the aims and objectives of this research, however, the level of response and representation of population groups was not satisfactory. The numbers of non-African deaths reported to have occurred in the previous 24 months were too few, which led to the African and non-African groups being

combined, after initially analysing the population groups separately because of expected different mortality patterns. Combining the two population groups, usually divided into African, Asian/Indian, Coloured and White population groups for analysis because of different mortality trends may also have affected the distribution of ages at death.

Weights were not incorporated into the research because they would make the procedure difficult in the handling of missing ages at death. However, unweighted data resulted in proportions of deaths that are below those of the vital registration for the age groups above 55 years. If the unweighted survey under represents the rural areas and a relatively higher proportion of older people die in the rural areas compared to younger people, this could produce the underestimation at the older ages observed.

To investigate the effect of weighting, the completed dataset created after multiple imputation had weights applied. Figure 5.1 shows that despite erratic distribution of the proportions for the younger ages, the distribution of ages above 55 years still show proportions that are lower than the vital registration, but less so, with an excess for the open interval 90 +.

Figure 5.1 Proportions of death by age: vital registration compared to estimates after multiple imputation (weighted)



The missing age pattern of ages at death data was investigated and it was concluded that data are missing completely at random because the percentages of missing data by relationship status did not show a particular pattern. However, adults are prone to remember the ages for children well, compared to the ages of old people. This is revealed in the NIDS dataset by the high number of implausible age differences between the deceased and the head of household in adult relationships. When the missing pattern is wrongly specified, likelihood based imputation procedures, such as multiple imputation yield wrong estimates.

To test the plausibility of the datasets created, external information is used, in particular, the vital statistics are used as the benchmark. Statistically, such a method of hypothesis testing may be regarded as informal testing because the similarities may be enhanced or distorted by the errors inherent in the dataset used as the benchmark (Abayomi *et al.*, 2008). With better data to compare with, marked differences of the model imputed data would provide evidence to investigate the imputation mechanism, the missing mechanism and the data, having full knowledge that the benchmark is correct. Imputation models and imputed datasets are usually compared using final estimates such as the efficiency of estimators, however, this is not done in this research because the objective was not to model the data but to generate data to replace the missing observations.

Comparing the hotdeck method and the multiple imputation method incorporating the relationship of the deceased to the head of household shows results that are quite similar. The differences being that the multiple imputation model increases variability in the data as seen by the proportional differences between the observed and complete data, while hotdecking produces distributions that are more peaked in the middle ages in comparison to the observed data. Examples of where peaked distributions after missing ages were replaced using a hotdeck, were observed by Dorrington *et al.* (2004) and Williams (1998) for the 2001 South African Census and 2000 American Census, respectively. The variability depicted by empirical density estimates after multiple imputation are distinctive, The imputations may be averaged by use of LOWESS curves, although, applying LOWESS curves would reduce the variance. Such density distributional changes are expected when values are missing at random. However, if the differences were striking, such as transforming from a bell-shaped density distribution to a table shaped density distribution would prompt investigation of the observed dataset, the missing mechanism and the assumptions underlying the predictive imputation model employed.

The research shows that using the variable, age of the head of household, and the relationship of the deceased to the head of household results in a complete dataset with proportional distributional characteristics comparable to the hotdeck method and registered ages at death. Statistics South Africa's hotdeck used five covariates and the Stata hotdeck used four covariates. This shows that from limited data, the multiple imputation method produces results that are similar to methods that use many variables which show that the multiple imputation method suggested is parsimonious with respect to these versions of hotdecking.

The Kolmogorov-Smirnov test, testing if the distribution of the observed values is equal to the distribution of the imputed values, showed that the distributions of data were different. The reason for the difference is that the imputation procedure was limited to relationships, and the percentages of the missing ages by the relationship status differed by a wide margin, with six of the relationships having at least 30 per cent missing and two relationships having less than 10 per cent missing. In addition, Table 3.3 shows that there was also a large difference between the numbers in the relationships, with the ages of deceased children of male headed households having 175 cases and deceased grandfathers having 6 cases.

Simple apportionment that allocates ages to the whole age range and simple apportionment that allocates ages to adult ages, above 20 years, produced completely different distributions from one another. When missing ages were allocated to the whole age range the proportional differences to the observed ages are equal to zero, revealing that the proportional distribution of the complete dataset is not affected by imputing missing ages. In comparison to hotdecking, the method produced a similar proportional distribution curve, slightly peaked in the middle ages. However, the small peak in the middle ages for the proportional distribution curve is inherent in the observed data, extended when hotdecking is used. Multiple imputation was able to reduce the peak in the middle ages. Simple apportionment that allocates ages to adult ages produces a proportional distribution that is skewed to the right, which is different from the distribution of observed data.

Future research might include investigating multiple imputation that incorporates the relationship of the deceased to the head of household variable to impute missing ages at death when observations are grouped by population group and sex of the deceased. Such grouping may reduce confounding elements observed when both sexes are analysed in the same group. An investigation of such magnitude requires a sample that is representative of the population groups and types of relationships in South Africa.

In terms of patterns of missing or mechanisms of missing ages of the deceased, it would also have been beneficial to investigate how the missing pattern for ages at death was related to the age of the heads of households. That is to see if old heads of household had more missing ages of the deceased than young heads of household. This is because multiple imputation was dependent on the age of the head of household and also on the relationship of the deceased to the head of household. Thus, there would be bias if the missing ages at death were only for the older heads of household compared to the younger heads of household by relationship status or without considering relationships.

In general, the impact of an imputation method is exposed when the percentage of missing values is high. The current environment where individuals are concerned with privacy and confidentiality as observed by Prewitt (2004), this may give rise to non-response in surveys or censuses. Therefore, tried and tested imputation methods are important to complete the records with missing data. Especially, methods that do not require extensive knowledge about the dataset and also do not require a large amount of observed data to replace missing values. However, as Dorrington *et al* (2004) posit imputation cannot be successful on poorly collected data.

References

Abayomi, K., Gelman, A. & Levy, M. 2008."Diagnostics for Multivariate Imputations", *Journal of the Royal Statistical Society. Series C (Applied Statistics),* **57**(3)**:**273-291.

Buuren, S. V. 2005.*What is multiple imputation?* http://web.inter.nl.net/users/S.van.Buuren/mi/. Accessed: 30 August 2010

Charlton, J. 2004."Editorial: Evaluating Automatic Edit and Imputation Methods, and the EUREDIT Project", *Journal of the Royal Statistical Society. Series A (Statistics in Society),* **167**(2)**:**199-207.

Cleveland, W. S. 1979."Robust Locally Weighted Regression and Smoothing Scatterplots", *Journal of the American Statistical Association,* **74**(368)**:**829-836.

Darikwa, T. B.2009."Estimating the level and trends of child mortality in South Africa, 1996-2006."Unpublished Masters thesis, Cape Town:University of Cape Town.

Dorrington, R., Moultrie, T. & Timaeus, I. M.2004.*Estimation of Mortality using the South African Census 2001 data.*Cape Town:University of Cape Town. http://www.commerce.uct.ac.za/Research_Units/CARE/Monographs/Monographs/Mono11.pdf.

Easterling, R. G. 1976."Goodness of Fit and Parameter Estimation", *Technometrics,* **18**(1)**:**1-9.

Faucett, C. L., Schenker, N. & Taylor, J. M. G. 2002."Survival Analysis Using Auxiliary Variables Via Multiple Imputation, with Application to AIDS Clinical Trial Data", *Biometrics,* **58**(1)**:**37-47.

Foote, K. A., Hill, K. H. & Martin, L. G. 1993. *Demographic Change in Sub-Saharan Africa.* Washington D.C.:National Academy Press.

Giles, P. 1988."A Model for Generalized Edit and Imputation of Survey Data", *The Canadian Journal of Statistics / La Revue Canadienne de Statistique,* **16**(1)**:**57-73.

Kaiser, J. 1983."The Effectiveness of Hotdeck Procedures in Small Samples,"Paper presented at the Annual meeting of the American Statistical Association,Toronto,17 August 1983.

Kirkman, T. W. 1996.*Statistics to use* http://www.physics.csbsju.edu/stat/. Accessed:10 September 2010

Kmetic, A., Joseph, L., Berger, C. & Tenenhouse, A. 2002."Multiple Imputation to Account for Missing Data in a Survey: Estimating the Prevalence of Osteoporosis", *Epidemiology,* **13**(4)**:**437-444.

Lepkowski, J. M., Landis, J. R. & Sharon, A. S. 1987."Strategies for the Analysis of Imputed Data from a Sample Survey: The National Medical Care Utilization and Expenditure Survey", *Medical Care,* **25**(8)**:**705-716.

Little, R. J. A. 1988."Missing-Data Adjustments in Large Surveys", *Journal of Business & Economic Statistics,* **6**(3)**:**287-296.

Mander, A. P. & Clayton, D. 2007. HOTDECK: Stata module to impute missing values using the hotdeck method. Boston, Boston College.

Manzari, A. 2004."Combining Editing and Imputation Methods: An Experimental Application on Population Census Data", *Journal of the Royal Statistical Society. Series A (Statistics in Society),* **167**(2)**:**295-307.

Moultrie, T. & Dorrington, R.2009.*Demography:Analysis of the NIDS Wave 1 Dataset.*Cape Town:University of Cape Town. http://www.nids.uct.ac.za/home/index.php?/Nids-Documentation/discussion-papers.html.

National Statistical Services 1997.*Statistical Clearing House-Basic Survey Design* http://www.nss.gov.au/nss/home.nsf/SurveyDesignDoc/5E4D15E56092F0E5CA2571AB00247A4D?OpenDocument. Accessed: 22 March 2010

Nordholt, E. S. 1998."Imputation: Methods, Simulation Experiments and Practical Examples", *International Statistical Review / Revue Internationale de Statistique,* **66**(2)**:**157-180.

Nordholt.E.S 1998."Imputation: Methods, Simulation experiments and Practical Examples", *International Statistical Review,* **66**(2)**:**157-180.

Parker, J. D. & Schenker, N. 2007."Multiple imputation for national public-use datasets and its possible apllication for gestational age in United States Natality files ", *Paediatic and Perinatal Epidemiology,* **2**(21)**:**97-105.

Paul, T. V. H. 2007."Regression with Missing Ys': An Improved Strategy for Analyzing Multiply Imputed Data", *Sociological Methodology,* **37**(1)**:**83-117.

Peugh, J. L. & Enders, C. K. 2004."Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement", *Review of Educational Research,* **74**(4)**:**525-556.

Prewitt, K. 2004."What If We Give a Census and No One Comes?" *Science,* **304**(5676)**:**1452-1453.

Rao, J. N. K. 1996."On Variance Estimation With Imputed Survey Data", *Journal of the American Statistical Association,* **91**(434)**:**499-506.

Refaat, M. 2007. *Data Preparation For Data Mining Using SAS.* San Francisco:Morgan Kaufmann Publishers.

Rockwell, R. C. 1975."An Investigation of Imputation and Differential Quality of Data in the 1970 Census", *Journal of the American Statistical Association,* **70**(349)**:**39-42.

Royston, P. 2004."Multiple Imputation of Missing Values ", *The Stata Journal,* **4**(3)**:**227-241.

Royston, P. 2005."Multiple Imputation of missing values: Update of Ice", *The Stata Journal,* **5**(4)**:**527-536.

Royston, P. 2007."ICE: Stata module for multiple imputation of missing values", *Stata Journal,* **7**(4)**:**445-464.

Rubin, D. B. 1976."Inference and Missing Data", *Biometrika,* **63**(3)**:**581-592.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys* New York:John Wiley & sons.

Rubin, D. B. & Schenker, N. 1986."Multiple Imputation for interval estimation from simple random samples with ignorable nonresponse", *Journal of the American Statistical Association,* **81**(394)**:**366-374.

Saari, D. G. 1978."Apportionment Methods and the House of Representatives ", *The American Mathematical Monthly,* **85**(10)**:**792-802.

Saldru2009.*National Income Dynamics Study User Document* Cape Town:University of Cape Town. http://www.nids.uct.ac.za/home/index.php?/Nids-Documentation/documents.html.

Sande, I. G. 1982."Imputation in Surveys: Coping with Reality", *The American Statistician,* **36**(3)**:**145-152.

Scheffer, J. 2002."Dealing with missing data", *Research Letters in the Information and Mathematical Sciences,* **3**(1)**:**153-160.

Statacorp. 2009. *Stata:Release 11. Statistical Software.* College Station, Texas:StataCorp LP.

Statistics South Africa2003a.*Census 2001:Computer Editing Specifications.*Pretoria:Statistics South Africa. http://www.statssa.gov.za/census01/html/editingspecs.pdf.

Statistics South Africa2003b.*Census 2001:Metadata.*Pretoria:Statistics South Africa. http://www.statssa.gov.za/census01/HTML/introduction.pdf.

United Nations 2000."Glossary of terms on statistical data editng,"Conference of European statisticians methodological material,Geneva,2000.

United Nations 2001. *Handbook on Population and Housing Census Editing.* New York:United Nations.

Williams, T. R.1998.*Imputing Person Age for the 2000 Census Short Form: A Model Based Approach.*Washington D.C:U.S.Bureau of the Census. http://www.amstat.org/sections/srms/proceedings/papers/1998_115.pdf.

Zang, B. & Walker, C. M. 2008."Impact of Missing Data on Person-Model Fit and Person Trait Estimation", *Applied Psychological Measurement* **32**(6)**:**466-479.

Appendices

Appendix A1 Stata do-file to merge mortality data and person data related to the head of household

```
clear
set mem 200m

cd d:
global IN "\CHNFAR(R&S)\research\NIDS\Stata_version10\Stata_version10"
global OUT "\CHNFAR(R&S)\research\NIDS\Stata_version10\Stata_version10\results"
global vIN "Anon_30Sep2009"
global vOUT "Anon_30Sep2009"
*-------------------------------------------------------------------------------
use "$IN\HouseholdQ_$vIN", clear

keep hhid  w1_h_mrt24mnth-w1_h_mrtacc5

egen num_miss=rowmiss( w1_h_mrtgen1-w1_h_mrtacc1)

bro if w1_h_mrt24mnth==1& num_miss==6

save "$OUT\HouseholdQ_$vOUT", replace
*-------------------------------------------------------------------------------
use "$IN\HouseholdRoster_$vIN", clear

keep hhid pid w1_r_gen w1_r_age w1_r_relhead

gen head=.
replace head=1 if w1_r_relhead==1
replace head=1 if w1_r_relhead==2
drop if  head!=1

save "$OUT\HouseholdRoster_$vOUT", replace
*-------------------------------------------------------------------------------
use "$IN\Adult_$vIN", clear

keep pid hhid w1_a_popgrp

ren w1_a_popgrp w1_popgrp

save "$OUT\Adult_$vOUT", replace
*-------------------------------------------------------------------------------
use "$IN\Child_$vIN", clear

keep pid hhid w1_c_popgrp

ren w1_c_popgrp w1_popgrp
```

```
save "$OUT\Child_$vOUT", replace
*-------------------------------------------------------------------------
use "$IN\Proxy_$vIN", clear

keep  pid hhid w1_p_popgrp

ren w1_p_popgrp w1_popgrp

save "$OUT\Proxy_$vOUT", replace
*-------------------------------------------------------------------------
use "$OUT\Adult_$vIN", clear

append using "$OUT\Proxy_$vIN" "$OUT\Child_$vOUT"

sort hhid
gen num = 1 if hhid !=hhid[_n-1]
replace num = num[_n-1] +1  if num !=1
drop if num!=1

save "$OUT\Popgroup_$vOUT", replace
*-------------------------------------------------------------------------
use "$OUT\HouseholdRoster_$vOUT", clear

merge 1:1 hhid using "$OUT\Popgroup_$vOUT"

drop pid head num _merge

save "$OUT\Heads_$v0UT", replace
*-------------------------------------------------------------------------
use "$OUT\Heads_$v0UT", clear

merge 1:1 hhid using "$OUT\HouseholdQ_$vOUT"

drop  num_miss _merge

reshape long w1_h_mrtgen@ w1_h_mrtr@ w1_h_mrtdod_m@ w1_h_mrtdod_y@
        w1_h_mrtage@  w1_h_mrtacc@, i( hhid) j(number)

egen num_miss=rowmiss(w1_h_mrtgen-w1_h_mrtacc)

drop if num_miss==6

save "$OUT\Mortality_$v0UT", replace
```

Appendix A2 Testing the equality of regression coefficients that are generated from two different regressions, estimated on two different samples.

The procedure taken in STATA to combine the deceased partners' regressions for African and Non-African female heads of households is illustrated.

```
. clear

. set obs 13
obs was 0, now 13

.
. gen x=invnormal(uniform())

.
. gen y=17.00593723+1.2985581*x+2*invnormal(uniform())

.
. gen d=0

.
. regress y x
```

Source	SS	df	MS	Number of obs =	13
				F(1, 11) =	21.76
Model	74.2965162	1	74.2965162	Prob > F =	0.0007
Residual	37.5597479	11	3.41452254	R-squared =	0.6642
				Adj R-squared =	0.6337
Total	111.856264	12	9.32135534	Root MSE =	1.8478

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	2.790676	.5982602	4.66	0.001	1.473914	4.107438
_cons	17.28948	.5161726	33.50	0.000	16.15339	18.42556

```
. save "C:\Users\Farai\Desktop\non_African.dta"
file C:\Users\Farai\Desktop\non_African.dta saved
```


Nested models are created, by combining the 2 regression models. In the process a dummy variable *d* and an interaction variable *w* are created. Testing whether the coefficient of the dummy variable and the interaction term are jointly zero is analogous to testing if the coefficients of the first regression are significantly different from the coefficients of the second regression as shown below. The conclusion is that the coefficients are not statistically different. Therefore the two samples may be combined to create one single model.

```
. set obs 52
obs was 0, now 52

.
. gen x=invnormal(uniform())

.
. gen y=21.41073309+1.157003348*x+2*invnormal(uniform())

.
. gen d=1

.
. regress y x
```

Source	SS	df	MS	Number of obs =	52
Model	75.8096898	1	75.8096898	F(1, 50) =	18.51
Residual	204.783297	50	4.09566594	Prob > F =	0.0001
				R-squared =	0.2702
				Adj R-squared =	0.2556
Total	280.592987	51	5.50182327	Root MSE =	2.0238

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	1.297174	.3015074	4.30	0.000	.6915781	1.902769
_cons	21.56005	.2834021	76.08	0.000	20.99082	22.12928

```
. save "C:\Users\Farai\Desktop\African.dta"
file C:\Users\Farai\Desktop\African.dta saved

. append using "C:\Users\Farai\Desktop\non_African.dta"

. gen w=x*d

. regress y x w d
```

Source	SS	df	MS	Number of obs =	65
Model	350.326934	3	116.775645	F(3, 61) =	29.39
Residual	242.343045	61	3.9728368	Prob > F =	0.0000
				R-squared =	0.5911
				Adj R-squared =	0.5710
Total	592.669979	64	9.26046843	Root MSE =	1.9932

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	2.790676	.6453206	4.32	0.000	1.500277	4.081074
w	-1.493502	.7103654	-2.10	0.040	-2.913966	-.0730387
d	4.27057	.622822	6.86	0.000	3.02516	5.515979
_cons	17.28948	.5567757	31.05	0.000	16.17613	18.40282

```
. test _b[d]=0

 ( 1)  d = 0

       F(  1,    61) =   47.02
            Prob > F =    0.0000

.

. test _b[d]=0, notest

 ( 1)  d = 0

. test _b[w]=0, accum

 ( 1)  d = 0
 ( 2)  w = 0

       F(  2,    61) =   27.91
            Prob > F =    0.0000
```

Appendix A3 Stata do for Multiple Imputation

creating the three imputed datasets for each relationship status

uvis regress varname HH , gen(varname_1) match

uvis regress varname HH , gen(varname_2) match

uvis regress varname HH , gen(varname_3) match

Appendix A4 Stata do for random selections between created datasets

creating one imputed data set for each relationship status

gen random_uniform=1+int((3-1+1)*runiform())

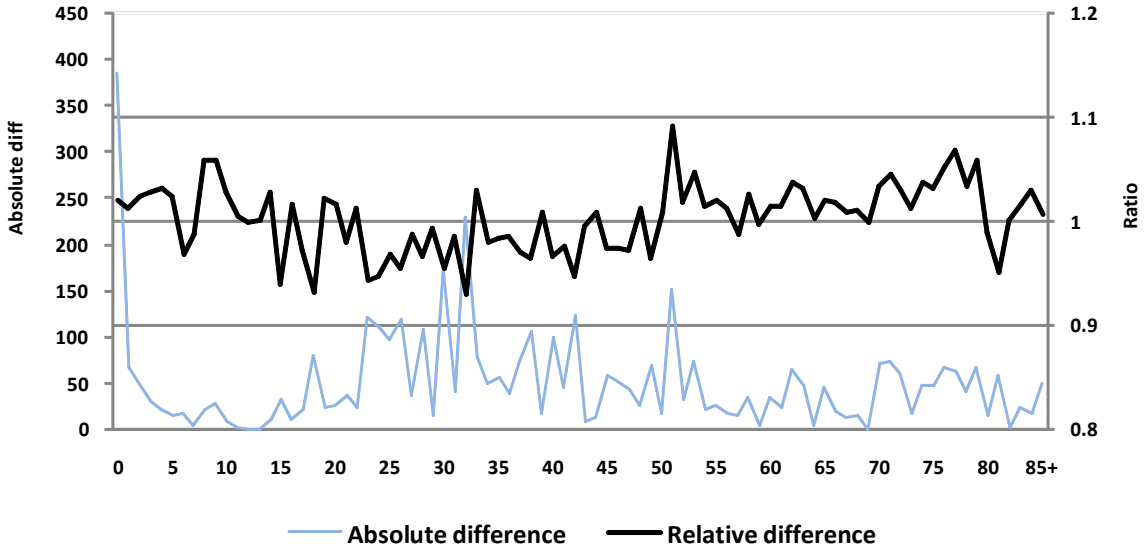
gen varname_11=cond(random_uniform==1, varname_1,.)

gen varname_22=cond(random_uniform==2, varname_2,.)

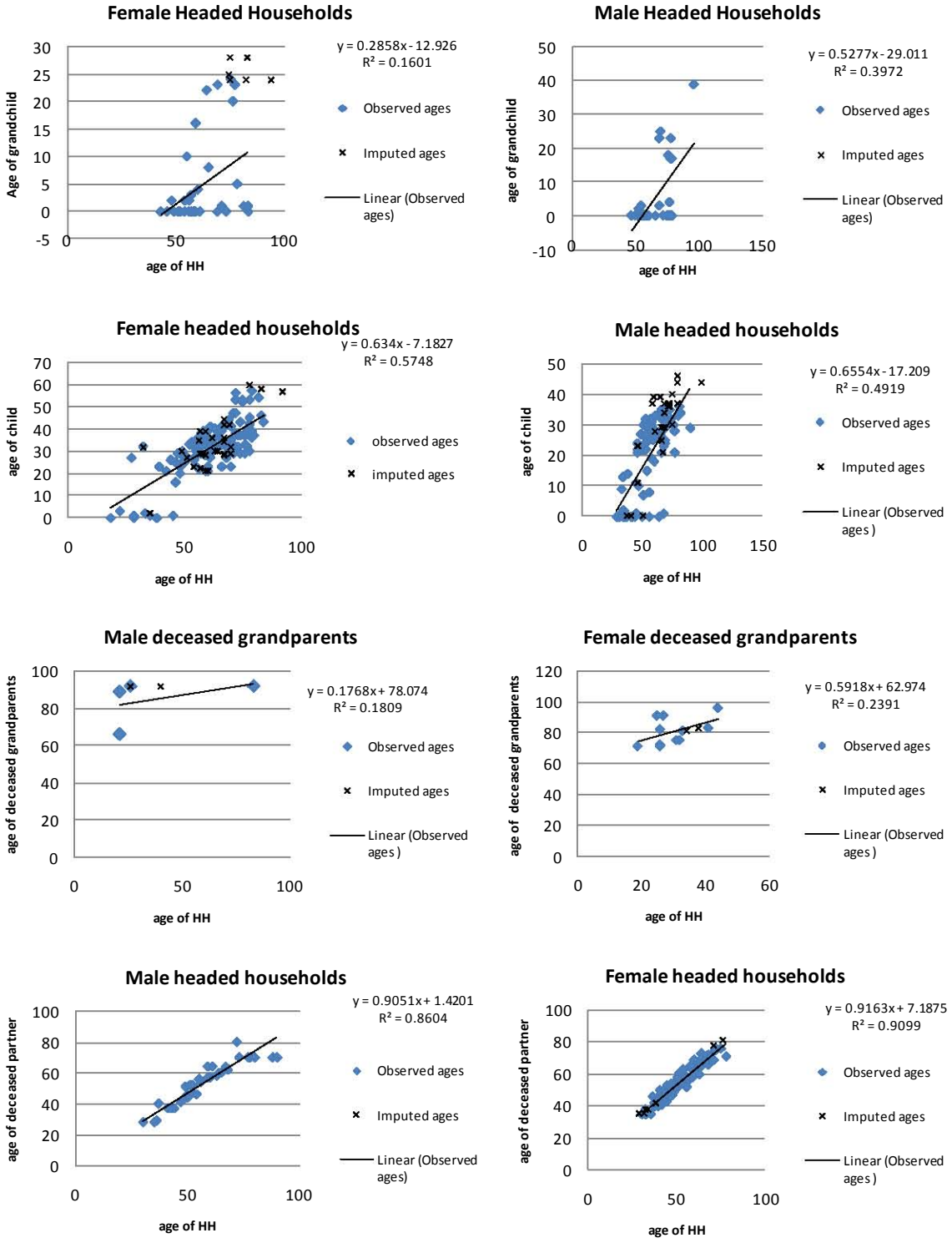
gen varname_33=cond(random_uniform==3, varname_3,.)

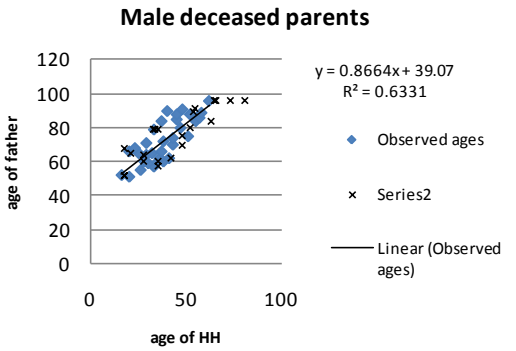
egen float imputed_var=rowfirst(varname_1 varname_2 varname_3)

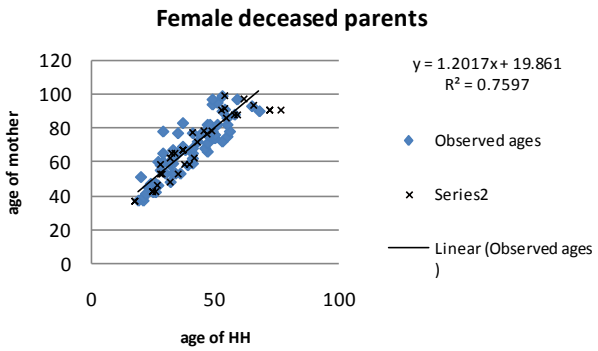
Appendix A5 Absolute differences and ratios by age between the two datasets after SOLAS hotdeck and Statistics South Africa's hotdeck.

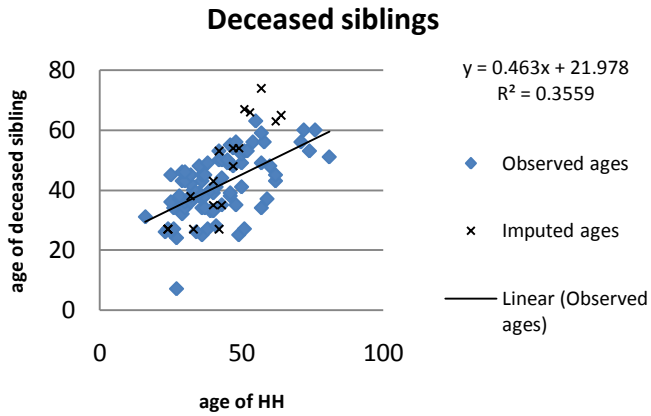


Appendix A6 Bivariate Scatter plots of Imputed data by relationship









Appendix A7 Stata Kolmogorov-Smirnov output

Stata do-file for Kolmogorov-Smirnov test and empirical density plots.

To conduct a Kolmogorov-Smirnov test

ksmirnov (varname), by(group) exact

To create an empirical density plot that starts at age 0.

kdensity (varname), nograph generate(x fx)

kdensity (varname) if group==0, nograph generate(fx0) at(x)

kdensity (varname) if group==1, nograph generate(fx1) at(x)

label var fx0 "observed ages"

label var fx1 "imputed ages"

drop if x>99&x<.

drop if x<0

line fx0 fx1 x, sort ytitle(Density)

Test for the equality between the observed data and the imputed data distributions after multiple imputation.

```
Two-sample Kolmogorov-Smirnov test for equality of distribution functions

 Smaller group       D       P-value     Exact
------------------------------------------------
 0:                  0.2093    0.000
 1:                  0.0000    1.000
 Combined K-S:       0.2093    0.000     0.000

Note: ties exist in combined dataset;
      there are 96 unique values out of 815 observations.
```

Test for the equality between the observed data and the imputed data distributions after using the Statistics South Africa imputation method.

```
Two-sample Kolmogorov-Smirnov test for equality of distribution functions

 Smaller group       D       P-value  Corrected
------------------------------------------------
 0:                   0.0113    0.963
 1:                  -0.0604    0.341
 Combined K-S:        0.0604    0.656      0.617

Note: ties exist in combined dataset;
      there are 97 unique values out of 951 observations.
```

Test for the equality between the observed data and the imputed data distributions after employing the Stata imputation method.

```
Two-sample Kolmogorov-Smirnov test for equality of distribution functions

 Smaller group       D        P-value  Corrected
 ------------------------------------------------
 0:                   0.0432   0.577
 1:                  -0.0527   0.442
 Combined K-S:        0.0527   0.809      0.779

Note: ties exist in combined dataset;
      there are 95 unique values out of 951 observations.
```

Test for the equality between the observed data and the imputed data distributions after employing simple apportionment of missing ages.

```
Two-sample Kolmogorov-Smirnov test for equality of distribution functions

 Smaller group       D       P-value      Exact
 ------------------------------------------------
 0:                   0.0000   1.000
 1:                  -0.0271   0.809
 Combined K-S:        0.0271   1.000      1.000

Note: ties exist in combined dataset;
      there are 97 unique values out of 940 observations.
```

Test for the equality between the observed data and the imputed data distributions after employing simple apportionment of missing ages.(20+)

```
Two-sample Kolmogorov-Smirnov test for equality of distribution functions

 Smaller group       D       P-value      Exact
 ------------------------------------------------
 0:                   0.1604   0.001
 1:                  -0.0052   0.992
 Combined K-S:        0.1604   0.001      0.001

Note: ties exist in combined dataset;
      there are 97 unique values out of 942 observations.
```

Appendix A8 Stata hotdeck command for hotdecking by population group

hotdeck w1_a_popgrp w1_h_mrtgenx w1_h_mrtdod_mx w1_h_mrtdod_yx w1_h_mrtagex using imputed_2, by(w1_a_popgrp) store impute(5) keep(hhid)

where w1_a_popgrp is the population group of the deceased

w1_h_mrtgenx is the gender of the deceased x^{th} person

w1_h_mrtdod_mx is the month of death for the deceased x^{th} person

w1_h_mrtdod_yx is the year of death for the deceased x^{th} person

w1_h_mrtagex is the age of the deceased x^{th} person