UNIVERSITY OF SOUTHAMPTON

THE USE OF CLASSIFICATION TREE TECHNIQUES FOR MISSING ITEM IMPUTATION

By

Dulce Maria Mesa-Avila

Doctor of Philosophy

Department of Social Statistics

Faculty of Social Sciences

July 2002

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL SCIENCES
SOCIAL STATISTICS

Doctor of Philosophy

THE USE OF CLASSIFICATION TREE TECHNIQUES FOR MISSING ITEM IMPUTATION

By Dulce Maria Mesa Avila

Censuses are the most important statistical demographic operation carried out by any country. The fundamental attribute regarding censuses is that they allow governments, interested organisations and researchers to handle key demographic information at any geographical level within a country. As any statistical collection processes, censuses are susceptible to what is technically known as "non-response". Non-response occurs when any investigated variable for any unit within the "universe of study" is missing in the final format for the analysis. Non-response can affect analysis, leading to erroneous or invalid findings and consequent decision-making.

This thesis compares different methods for imputing item non-response present in census information based on classification. The strategy for carrying out the imputation is divided in two steps. First, the data set is classified using a Tree-Based Technique, and second, the imputation is made using some of the known imputation methods.

The "Classification and Regression Tree" (CART) technique used for tree-based modelling is basically a set of classification rules (recursive binary segmentation) that partition the data set into mutually exhaustive and non-overlapping subsets (terminal nodes) based on the values of a group of explanatory variables. These subsets are expected to be internally more homogeneous with respect to the response variable (the variable for which the tree is generated) than the whole database. Once the classification is made, each imputation method is applied independently within each terminal node. Three common imputation methods for categorical data are used.
The combination of classification and imputation makes possible the assessment of the following aspects: 1) the effect of using this classification technique on the imputation results (including the use of different tree-sizes), and 2) the accuracy of the different imputation methods based on this classification technique.

The analysis was carried out for two different settings: the univariate case where a single variable is imputed, and the multivariate case where a composite variable is imputed. A composite variable is defined by the cross-classification of two or more single variables. The use of the composite variable allows for the imputation of two or more single variables at the same time.

The preservation of joint and individual marginal distributions as well as the preservation of individual values are evaluated. Graphs and tests for those comparisons are presented. Additionally, assessment of biases and variances, as well as variance estimation in some cases, are presented.

The simulation was made using a subset of UK 1991 Census information. Only categorical variables related to persons (except age, which was converted to categorical) were used for the analysis. After deleting the records with missing information from the original database, artificial holes were created using the real pattern of missing information present in the original database. This makes possible the measurement of the accuracy of the imputation by comparing the real values and the imputed values.

Some general conclusions are obtained from the simulations: 1) the use of the classification tree as a method for creating imputation cells before the imputation is carried out does improve the imputation results, although the size of the tree does not have a major impact on the results. 2) most of the imputation procedures used in the simulation produce unbiased estimates for the total and for the variance, additionally, they have a very high values for the coverage as well as low values for the relative Mean Square Error, 3) in general, the best performing method is the Frequency Distribution method (even when compared with Sequential Hot Deck imputation).

TABLE OF CONTENTS

CHAPTER 1: *INTRODUCTION*

CHAPTER 2: *TREE-BASED CLASSIFICATION*

CHAPTER 3: *UNIVARIATE CASE - THEORETICAL FRAMEWORK*

CHAPTER 4: *UNIVARIATE CASE - SIMULATION*

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEGMENT

I would like to thank my supervisor Prof. Chris Skinner for the quality of his supervision and his criticism and encouragement throughout my study-period. I would also like to thank Prof. Ray Chambers, whose input and support has been more than invaluable.

Thanks are also due to all the member of the Social Statistics Department for their unconditional cooperation and assistance, particularly to Anne Owens and Sandy Mackinnon.

I would also like to express my gratitude to all my colleagues and friends. Among them, Maria Norton, Penny Tsai, Tiziana Leone, Priscilla Akwara and Juliet MacEachran.

Further thanks are due to many people in Venezuela for their support and assistance. I am especially grateful to Emiro Molina and his wife Julie for their encouragement and friendship. Thanks are also due to Miguel Bolívar, Jose Domingo Mujica, Alberto Camardiel, Guillermo Ramírez and Omaira Santoyo.

I am forever indebted to my parents Dulce and Luis and to my brother Luis Roberto for all their support and encouragement during all these years. Finally, no words can express my gratitude to my husband Felix Leonardo for his unconditional support and understanding, encouragement and love throughout all this process.

The data used was made available by the Census office of the Office for National Statistics in the United Kingdom.

This Ph.D. was part-fee supported by the ORS Award Scheme (United Kingdom Scholarship for International Students) during three years.

CHAPTER 1

INTRODUCTION

1.1. CENSUS DATA AND NONRESPONSE

Census data are the most important source of statistical information for a country. This is the basis of planning and decision-making by governmental offices. In addition, as an important attribute, census information may be used in the sampling frame for many surveys carried out by either government or any other official institute.

The collection process of census information requires much effort and some complex procedures. These two factors, together with the magnitude of the data collected, make census information susceptible to missing values. The missing information problem introduces difficulties to the analysis.

Several factors can lead to missing information in a census (or survey in general). Noncoverage, measurement errors and nonresponse are the most common sources of missing information. No one solution exists to remedy all of these issues. Different problems require different solutions. Despite this, some techniques are valid for more than one problem; for example, imputation could be used not only for filling in missing values due to nonresponse, but also for replacing items with measurement errors as these can be treated as missing values in the final data once they have been detected by the editing process.

Even though we have said that missing information may arise from different sources, nonresponse will be the main aspect considered in this thesis.

Nonresponse is caused by the incapacity to obtain complete measurements for any of the units or variables in a census (survey). There are two different kinds of nonresponse: *Unit Nonresponse* and *Item Nonresponse*. Unit nonresponse occurs when a unit fails to participate in the study. In this case, none of the variables or information for that unit is collected. In contrast, item nonresponse occurs when information for one or more items in an accessible unit is not available. This means that only a part of the information for that unit is collected. Different factors can generate nonresponse. Unit nonresponse can be caused by temporary or permanent unavailability of the unit; because the unit refused to participate, or it might be unable to answer the questions. On the other hand, causes for item nonresponse include interviewee refusal or inability to answer a specific question; perhaps the interviewer omits to ask the question or fails to record the answer for a specific item. Additionally, after the editing process, records with invalid responses that cannot be fixed by this process are also considered within the pool of nonresponse in order to be imputed.

This thesis considers the case of item nonresponse as the main problem to be investigated.

In order to illustrate the item nonresponse problem, Tables 1.1.1 show an example of the percentages of item missing presented in a single county of the UK for 1991 census.

Tables 1.1.1

Missing information for a single County in the UK

Table A
Household

Variable	% missing
Bath and shower facility	0.12
Brick building	1.55
Building type	3.82
Number of cars owned by the household	2.06
Central heating facility	0.48
Number of rooms occupied by household	25.59
Ownership/rental status	1.21
Type of accommodation	5.00
Toilet facility	0.13

Table B
Individuals

Variable	% missing
Age	7.20
Primary activity last week	5.83
Country of birth	7.26
Ethnicity	16.24
Long term illness	13.37
Marital status	7.05
Sex	4.93
Ability with welsh	13.39

These tables show an example of the percentage of item missing for single variables, however, census information commonly has different missing items for the same person. That is, combination of missing values between these variables may be also present. Therefore, we can see that it is not a straightforward problem given the amount of missing information present and the complexity of the missing pattern (all possible combinations missing).

There are several consequences of nonresponse for the analysis when the loss of information is too high within persons or households. Reduction of the number of units composes one of these consequences, generating some problems, which will be explained later in this chapter, for analysing the results (including biased estimates), especially when the size of the population is not too large (Lessler and Kalsbeek, 1992; Sande 1982; Madow, et al 1983). Also, when more than one variable is missing at the same time, the size of the population may change from one variable to another, making estimation, computation, and comparisons more difficult.

Another crucial consequence of nonresponse is the possible presence of bias. This is an important aspect to be taken into account since it could generate a completely misinformed analysis. Bias is usually considered as the main measure of the nonresponse impact on the results (Lessler and Kalsbeek, 1992). The bias quantifies the difference between the expected value of an estimator over all possible samples and the population value.

Because of the progressive increase of the nonresponse problem over the years, many solutions have been developed recently to address this problem (Lessler and Kalsbeek, 1992; Kalton and Kasprzyk 1982; Sande 1982; Madow, et al 1983). One of the most important attempts for reducing this problem is an endeavour to decrease or at least control the nonresponse rates present by making some additional effort at the moment of interview or post-interview. However, this is not enough to eliminate the problem. Sometimes, it is more complicated or too expensive to go back to the interviewee in order to get the full answer; maybe the interviewee simply does not know or is unwilling to give an answer. Consequently, because of the incapacity of researchers to obtain complete data, different compensating procedures have been developed to confront the problem. These compensating procedures are commonly made by using weighting or imputation methods. In general, weighting procedures are used in the case of unit nonresponse or noncoverage, while imputation procedures are used in presence of item nonresponse (Lessler and Kalsbeek, 1992, Kalton, 1983, Madow et al 1983). However, in the case of census data, post-enumeration surveys are generally used to deal with the noncoverage problem (e.g. Ericksen, Kadane and Turkey, 1989; Breiman,1994; Kearney, A. Ikeda, M. 1999).

Our main concern is the item nonresponse problem in census data. Therefore, the objective of this thesis is to consider compensating procedures for this specific problem.

1.2 COMPENSATING FOR ITEM NONRESPONSE

Given that item nonresponse is usually present in most data, different solutions can be used in order to deal with this problem. One solution could be to analyse the data by either using the available cases, which uses all the available values or deleting the cases with missing values in order to use complete cases (Little and Rubin, 1987).

In the first case, available cases, there are several complications, starting with the fact that sample sizes change from variable to variable creating complications for making tabulations including many variables or when comparisons across variables are made. Also, some procedures for analysing the data, as well as some computational programs, make use of only complete cases. Additionally, unless the mechanism generating the missing information is completely at random, the introduction of the bias could also be a considerable problem.

In the second case, complete cases, there are both advantages and disadvantages. The use of a common sample (only complete cases) and therefore the use of standard methods make the analysis simpler and easier. However, there could be a very high loss of information when discarding incomplete cases, as well as a loss in the sample size. In addition, if the information is not missing completely at random, the introduction of bias in the results may represent an important aspect to consider when using the resultant information.

Another possible option could be to separate the units containing missing information to a different category, which can be called "unknowns". However, this procedure is still ignoring information available for other variables within the same unit. In this case, analyst will generally refer to the unknowns as a category without being able to use the micro data (Sande, 1982).

Despite all the options mentioned before, it can be seen from the description of the methods mentioned above that these have been insufficient in solving the item nonresponse problem. As a result of this, the use of compensating methods in the presence of missing information has increased over the last several years, making simpler analysis possible. In fact, the multivariate nature of the information collected in census, where all the variables can be subject to nonresponse, makes the use of compensating procedures for item nonresponse more necessary and useful.

As mentioned before, item nonresponse problem is generally solved by the use of imputation procedures.

In general, the use of imputation procedures implies certain pros and cons (Kalton, 1983; Kalton and Kasprzyk, 1986; Sande 1982; Lessler and Kalsbeek, 1992).

Some important advantages to be mentioned include:

1. As any compensating procedure, imputation aims to reduce the biases in the estimates arising from nonresponse;
2. Imputation makes the analysis easier and the results simpler to present, i.e. no complex procedures for analysing incomplete data are required;
3. Results from different analyses are bound to be consistent;
4. Imputation assigns values at the micro-level, which allows for a more complex analysis (taking into account the correspondent considerations or restrictions)

Important disadvantages of using imputation methods constitute the following:

1. Less bias is not guaranteed after the imputation has been done. In fact, bias can be greater (depending on the suitability of the assumptions built into the method used);
2. Bias of univariate statistics can be reduced while the relationship between variables could be distorted;

3. The data could be used as a complete set overstating the precision of the estimates;

There are also a number of problems to be dealt with when using an imputation procedure. Sande (1982) describes some of them in the following manner:

1. The close relationship between editing and imputation. It is not easy to decide which record(s) (or item(s)) has to be imputed when an edit fails. Additionally, the imputed records must satisfy the edit constraints in order to produce consistent data. Fellegi and Holt (1976) propose a methodology for dealing with this problem, which specifies that the imputation must be done by changing as few items as possible (among other aspects) (see *Section 1.7*).
2. Different records can have different patterns of missing information. This makes the decision regarding selection of an imputation procedure more difficult.
3. The time constraints constitute very important factors to be taken into account. Normally, there is no time for testing with the data until it is ready.
4. The use of imputation does not guarantee better results compared to using classical estimation techniques for incomplete data. In fact, it could sometimes be considered worse to use imputation.
5. Estimates from imputed data could be less reliable than when complete data are used. Normally, the estimation of variances is inadequate, as they do not include error arising from imputation.
6. The ethical problem of giving out the micro-data. Alternatives such as identifying the imputed values or giving the edited but non-imputed data are options that have to be decided upon.

1.3 NOTATION

Before starting on a description of imputation methods, let us begin with some notation that will be employed throughout this thesis.

Let U be a finite population of N units $U=\{U_i; i=1,2,...,N\}$. Let $\mathbf{Y}=(y_i)$ be a $(Nx1)-vector$ of response variable, where y_i represents the $i\,th$ element and let $\mathbf{X}=(x_{ik})$ be a $(NxK)-matrix$ of auxiliary variables with x_{ik} as the $k\,th$ variable for $i\,th$ the element. $\mathbf{X}$ can be represented as $\mathbf{X}=(\vec{X}_1,\vec{X}_2,...,\vec{X}_k,...,\vec{X}_K)$, where $\vec{X}_k=(x_{1k},x_{2k},...,x_{Nk})^t$ is a vector of N values x_{ik}.

Given that the aim of this research is to present an alternative solution for the missing information problem in census data, sampling is not considered in this thesis. That is, all the units in the population are included in the study.

Assuming that variable y_i are subject to nonresponse and x_{ij} are fully observed, we also define $\mathbf{R}=(r_i)$ as $(Nx1)-vector$ of indicator variables for $\mathbf{Y}$, which identifies whether or not y_i is missing. That is, $r_i=\begin{cases}1 & if\ y_i\ is\ observed\\ 0 & otherwise\end{cases}$.

Hence, the population can be represented as follows:

$$\begin{array}{cccccc}
x_{11} & \cdots & x_{1k} & \cdots & x_{1K} & y_1 \\
. & & . & . & & . \\
. & \cdots & . & . & \cdots & . \\
. & & . & . & & . \\
x_{m1} & \cdots & x_{mk} & \cdots & x_{mK} & y_m \\
x_{m+1,1} & \cdots & x_{m+1,k} & \cdots & x_{m+1,K} & 0 \\
. & & . & . & & . \\
. & \cdots & . & . & \cdots & . \\
. & & . & . & & . \\
x_{N1} & \cdots & x_{Nk} & \cdots & x_{NK} & 0
\end{array}$$

where m is the number of records for which $\mathbf{Y}$ is observed (measured) and the zeros represent the missing values. That is, we take, without loss of generality, $r_1=r_2=\ldots=r_m=1$ and $r_{m+1}=r_{m+2}=\ldots=r_N=0$.

This case corresponds to the univariate case in which only one variable is subject to nonresponse. However, this can be extended to the case in which many variables can be subject to nonresponse at the same time. In this case, $\mathbf{Y}$ and $\mathbf{R}$ will become a matrices of variables and it is explained in *Chapter 5*.

1.4 MISSING DATA MECHANISMS

The process by which the missing data are generated represents an important aspect when choosing a compensation procedure. Little and Rubin (1987) distinguish *ignorable and non-ignorable* mechanisms. An ignorable missing data mechanism is such that the missing values do not depend on the variable which is missing. On the other hand, in the case of the non-ignorable mechanisms, the missing information depends on the values of the absent variables.

There are two important ideas related to the concept of an ignorable missing data mechanism. First, the missing data are *Missing at Random* (MAR) when the probability of the variable being absent does not depend on the value of this variable conditional on observed information; and second, *Missing Completely at Random* (MCAR) when this probability of response does not depend either on the value of the missing variable or on the rest of the variables. In these two cases, the missing data mechanism is ignorable.
When missingness depends on the values of the missing variable and possibly on the rest of the variables, the missing data mechanism is non-ignorable.

Since most of the common imputation methods make assumptions about the probability of nonresponse, it is important to understand the missing data process in order to carry out imputation procedures. Therefore, these concepts will be explained in a more formal way hereafter.

As assumed above, suppose that $\mathbf{Y}$ is subject to nonresponse and $\mathbf{X}$ is fully observed. Let $\mathbf{R}$ be the response indicator for $\mathbf{Y}$.
Since the model treats r_i as random variables, let us consider the model where y_i, x_{ik} and r_i are all random variables. Then, we write the joint distribution as $f(\mathbf{Y},\mathbf{X},\mathbf{R})$.

When $\mathbf{R}$ is independent of $\mathbf{Y}$ given $\mathbf{X}$, that is $f(y_i \mid x_{ik}, r_i = 1) = f(y_i \mid x_{ik}, r_i = 0)$ the data is called *Missing at Random*, MAR, where $f(\mathbf{Y} \mid \mathbf{X},\mathbf{R})$ denotes the probability function of $\mathbf{Y}$ given $\mathbf{X}$ and $\mathbf{R}$.
When $\mathbf{R}$ is independent of $\mathbf{Y}$, that is $f(y_i \mid r_i = 1) = f(y_i \mid r_i = 0)$ the data are *Missing Completely at Random*, MCAR, where $f(\mathbf{Y} \mid \mathbf{R})$ denotes the probability function of $\mathbf{Y}$ given $\mathbf{R}$.
In these cases, the process that generates the missing data mechanism is ignorable as the missing data mechanism is such that the missing values do not depend on the variable which is missing.

1.5 IMPUTATION

Imputation is the process through which individual missing items are given a value in order to produce complete data (i.e. "imputation 'completes' incomplete responses" (Sande, 1982)). The information used to produce the imputed values normally comes from the respondents. The broad idea of imputation methods is to pick a replacement value that is as similar as possible to the missing item (Lessler and Kalsbeek, 1992).

Imputation procedures can be classified in different ways. A common way for grouping the imputation methods indicates a division into three very general groups (GSS Methodology Series, 1996):

Deductive methods, in which the values are deduced from known information, either from complete records or other available information, with certainty or high probability. This depends on some redundancy in the information collected; for example, if a member of a family if under 16, deduce marital status as "single".

Deterministic methods, when, under the same conditions, repeated imputations produce the same answers. Examples of this kind of imputation are mean (or mode) imputation, regression imputation, and nearest neighbour imputation.

Stochastic methods, when repeated imputations made under the same conditions can produce different results. This indicates that there is a random element included. Examples of this are imputing from randomly selected cases and regression imputation with a random term.

Another way to classify imputation methods depends upon the use of internal or external sources of information (Lessler and Kalsbeek, 1992; Kalton, 1983).

Cold-Deck procedures imply the use of external information (different from cases in the survey or census) for the imputation process. This makes use of information from different sources like, for example, past data sets from the same population. This method has the disadvantage of a potential lack of comparability between past and present values, which can be a problem when imputing (e.g. the use of different procedures for collecting the data or different definitions for a variable).

On the other hand, *Hot-Deck procedures* make use of the data available in the survey or census in order to create an imputation. Since the term Hot-Deck determines only whether the imputations are derived (or not) from the same data set, different ways of selecting the imputations can be used such as imputing from a randomly selected cases or nearest neighbour imputation.

Before embarking upon a description of the different imputation procedures, it is important to define some common concepts used in the area.

A Donor is the record from which the value to be assigned to the missing item is normally taken. The records with the missing items (for which the imputation is done) are called *Recipients*. It is important to point out that not all the imputation methods assign values to recipients from a donor, e.g. mean imputation.

Imputation is the value used in order to fill in the missing item. In the case of donor imputation this value comes from the same variable being imputed but from a complete case.

Auxiliary variables (also called *control variables, matching variables* or *assignment variables*) are those related to the variable with missing values. These are available for respondents and non-respondents. They are not only used for defining imputation cells, but also for defining regression models for imputing and quantifying how close donors and recipients are.

Imputation classes define partitions of the population made according to similarities generally based on the values of a set of auxiliary variables.

In order to describe a general imputation method, let $\hat{y}_i$ be the imputed value, and $\bar{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})$ a K-dimensional vector of auxiliary variables for the ith unit with actual value y_i. A wide class of imputation methods can be written in the following way

$$\hat{y}_i = g(\bar{x}_i) + e_i,$$

where $g(\cdot)$ is a function of the auxiliary variables, and e_i are specified residuals. In this case, the specification of the form of $g(\cdot)$ and whether the imputation is fixed or random (depending on the use of e_i) allow for the making of a distinction between the different imputation methods (Lessler and Kalsbeek, 1992).

In the case of deterministic imputation, $e_i = 0$, different specifications of $g(\cdot)$ can be written, for example,

$$\hat{y}_i = g(\bar{x}_i) = x_{i3}$$

or

$$\hat{y}_i = g(\bar{x}_i) = x_{i1}x_{i3} - x_{i4}.$$

In the case of linear regression imputation, the function of the auxiliary variables can be written as

$$g(\bar{x}_i) = b_0 + \sum_{k=1}^{K} b_k x_{ik}$$

where b_0 *and* b_k's are estimated by standard methods such as least squares or maximum likelihood. In the case of a categorical response variable, the regression can be done by using logistic or log linear models.

In the case of stochastic imputation, the linear regression function can be written as

$$g(\bar{x}_i) = b_0 + \sum_{k=1}^{K} b_k x_{ik} + e_i$$

in which a random term is used.

In the case of mean imputation, the definition of the imputed value has the form

$$\hat{y}_i = \bar{y}_o$$

where $\bar{y}_o$ represents the mean of the observed values for the variable (y_i) to be imputed. Here, the imputation does not depend on the auxiliary variables. However, the method can be generalised by taking the mean within imputation classes, which are defined by the x_{ik}. This is a deterministic method, which does not make use of any random term.

In the case of categorical data, $\hat{y}_i$ may take the value of the modal category and moreover, if the imputation is carried out within imputation classes, $\hat{y}_i$ may take the value of the modal category of the specific class to which i belongs.

In the case of nearest neighbour imputation, the imputed value is obtained from a donor which is selected according to a function of distance, which can be defined in many different ways. For example,

$$\hat{y}_i = y_{i'}$$

where $y_{i'}$ satisfies $\min_{i'}[d_{ii'}]$, with $d_{ii'} = \sum_{i'}^{K} I(x_{ik} \neq x_{i'k})$.

Some imputation methods are suitable for categorical variables while others are suitable for continuous variables. There are some cases in which both kinds of variables can be used and on some occasions, more than one method can be combined in order to determine a final strategy for imputing.

Since this research is mainly concerned with the missing information problem in population census data, the kind of variables used are principally categorical variables. Therefore, our interest will be imputation procedures for this particular kind of data. However, it is important to point out that the method proposed in this thesis can be used when treating not only categorical variables but also continuous variables or a mixture of them in any of the cases, independent or dependent variables, by making certain adjustments especially to the imputation methods.

An important method commonly used for imputation, especially in the case of census data, is what is traditionally called hot deck imputation. The traditionally called hot deck imputation is basically a sequential procedure in which given a set of imputation classes, within each class the records are treated in a sequential way. If a record has a response in the y_i variable, this value is stored replacing the previous one in order to be used for imputation. If a record has a missing value in the y_i variable, the value currently stored is assigned to that missing item. The starting value within each imputation class is normally assigned from previous surveys (census).

One of the most important advantages of using a hot-deck procedure is the use of information from the same census (survey), which can help maintaining relationships between variables while completing the missing information. Also, utilising information from the same investigation guarantees the use of the same theoretical context in terms of definitions and concepts used.

If the method traditionally called hot-deck is used, where the sequential procedure is involved, then, depending on the way in which the file is ordered, an additional degree of matching is introduced (Kalton, 1983). However, an important disadvantage of this procedure is that it may easily lead to the situation of multiple use of donors, which can contribute to a lowering of the precision of estimates and underestimation of the variances in surveys.

1.6 EVALUATION OF IMPUTATION

In order to choose an imputation procedure it is necessary to evaluate its performance. This section considers some ways in which this may be done.

The most common aspects to be taken into account when choosing the imputation procedure are (Lessler and Kalsbeek, 1992):

1. The statistical repercussion of the method on the estimates. It would be desirable to find a method that allows for doing the statistical inference intended while minimising the effect of the nonresponse on that inference.

2. The effect of the compensation procedure on the relationship between the variables. Cross-tabulations, regressions and other analyses for investigating relationships between variables can be affected by the method chosen.
3. The availability of the auxiliary data, if required. Some of the methods may require auxiliary information, which must be available to allow for the use of the compensation procedure selected.
4. Not only statistical effectiveness has to be considered, but also the practical implications. Sometimes, the compensation procedure is statistically adequate but extremely difficult (or even impossible) to implement because of the practical requirements. A compromise between statistical and practical effectiveness is important.
5. A review of recent comparative studies can also be useful in the selection of compensation procedure. The comparison can be made in the use of biases and variances (analytical comparison) or comparing real vs. imputed values (empirical comparison).

All these are important aspects to be taken into account when a procedure for imputing is to be chosen. However, many other factors can influence the decision. Despite knowledge about a large number of aspects, there are no rules for combining all of them in order to dictate how to establish the appropriate procedure for imputing. Nevertheless, the imputation procedure will be more accurate when taking into account as much information as possible within the decision process.

Some other criteria are also useful in choosing an imputation procedure as explained hereafter.
Before defining some procedures for this task, let us introduce some useful notation.
Let θ be the population parameter and $\hat{\theta}$ its estimator based upon the imputed dataset. Let $SE\left(\hat{\theta}\right)$ be the standard error for $\hat{\theta}$ and $\hat{\theta} \pm 2SE(\hat{\theta})$ its 95% confidence interval.

The first measure of performance is bias. A good imputation method requires low bias, which means $E(\hat{\theta})$ close to θ . As the bias itself may be difficult to interpret, there are two ways of looking at it in order to get alternative information for the evaluation of the imputation performance.

First, relative bias can be defined in the following way, $\frac{E(\hat{\theta})-\theta}{\theta}$. This can also be expressed as a percentage of θ . That is $\frac{E(\hat{\theta})-\theta}{\theta}*100$. These values are easier to interpret since they are relative measures.

A second way of using the bias for evaluation proposes is using a standardised bias, as a percentage of standard error. This is $\frac{E(\hat{\theta})-\theta}{SE(\hat{\theta})}*100$. "Once this exceeds 30-40% it starts to adversely affect coverage of confidence intervals" (Schafer 2001).

The second measure of performance is the variance. A good imputation method also requires low variances for $\hat{\theta}$, $V(\hat{\theta})$. However, low variance does not help very much when $\hat{\theta}$ is biased.

Combining the two aspects mentioned before provides a measure of the accuracy. That is, $MSE(\hat{\theta})=E\left[\left[\hat{\theta}-\theta\right]^2\right]=V\left(\hat{\theta}\right)+\left[E\left(\hat{\theta}\right)-\theta\right]^2$. High values of $MSE(\hat{\theta})$ are not desirable since they imply either big variance or big bias (or both).

Confidence intervals for the estimator of the parameters, $\hat{\theta}\pm 2SE(\hat{\theta})$, can also be used for evaluating the imputation procedure. They should be as narrow as possible, and include the true value of θ the specified proportion (e.g. 95%) of tries.

In this thesis, different ways of assessing the imputation procedures are used. One of them is the assessment of the properties of the different method used, including biases and variances. Additionally, the use of graphical methods and statistical tests for comparing distributions are also used in order to verify the validity of the results.

1.7 SOME IMPUTATION EXPERIENCES WITH CENSUS DATA

The main focus of this thesis is the use of imputation for missing information in population census data. Different approaches to imputation have been used by different organisations in charge of the census programs. Most offices for statistics use similar methods for imputing demographic categorical data, which is basically the kind of data obtained from a population census. For example, most offices use decision tables, look-up tables or hot-deck methods such as sequential hot deck, fixed-cell or nearest neighbour methods. In many cases, the use of these methods arises from a lack of financial or technical resources. Additionally, simplicity and time-saving are attributes highly important for governmental offices when treating census data given the amount of information involved and the urgency of the results.

It can also be said that the solution to the missing information problem is reduced, in some countries, to the use of editing and coding systems rather than imputation systems.

Editing is defined by Granquist (1997) as the procedure for identifying, by means of edit rules, and for adjusting, manually or automatically, errors resulting from data collection or data processing.

Granquist (1997) specifies in his article that there are three roles editing has, which are mentioned here in priority order:

- Identify and collect data on problem areas, and error causes in data collection and processing, producing the basics for the future improvement of the survey vehicle
- Provide information about the quality of the data
- Identify and handle concrete important errors and outliers in individual data

However, many countries use editing as a tool for cleaning up the data in order to have valid information. Some countries integrate the editing process with the data entry procedure such that the data is clean i.e. it passes all the edits, after the data entry routine is done.

It is important to mention that the editing process could be seen as imputation when codes are changed due to inconsistencies. However, the term imputation in this work refers to the use of any of the procedures known in the literature as imputation method.

Fellegi and Holt (1976) refer to the relationship between editing and imputation and the importance of the creation of an edit and imputation system which allows for the following:

- The data should be so that satisfy the edit rules by changing as fewer items as possible. That is, maintaining as much original information as possible.
- Imputation rules should be derived automatically from the edit rules in order to ensure the validity of the imputed records. That is, imputed records will continue to pass edit rules.
- The imputed data should be such that the individual marginal and joint distributions are maintained as far as possible.

Fellegi and Holt refer in this paper to topics like the application of logical edits to a record, derivation of a complete set of logical edits, derivation of a complete set of arithmetic edits, identification of the minimal set of fields for imputation and some procedures for imputation.

Therefore, we can see the close relationship between editing and imputation, however, this thesis is concern with the imputation aspect rather than the editing aspect of the census data.

Some examples of systems for imputation with census data developed by statistical offices in the recent years are NIM, EDIS and SCIA.

NIM (New Imputation Methodology) is a system developed by Statistic Canada used for population census data (Bankier, 1999; Poirier,1999; Bankier, Lachance & Poirier, 1999; Hill, 1976; Bankier, Houle, Luc & Newcombe, 1998). This system is essentially based on the nearest neighbour methodology, selecting the donor randomly from the pool of possible donors based on the minimum distance. The "feasible" donors are those that allow the recipient to pass the edits. This also introduces the Fellegi and Holt idea of minimum change, which comprises the use of as much of the information present in the data as possible (i.e. to change as few items as possible in a record) as explained before in this section. This system was used in the Canadian Census in 1996 and it has been reviewed during recent years achieving some improvements.

DIS (Donor Imputation System) is a system developed by the Office for National Statistics in the U.K (Anderson, F. and Whitfield K. 2000; Vickers, 1999; Vickers and Mohammed, 1998; Richards, 1999). This system includes the editing and imputation processes of the census and it is also based in the nearest neighbour methodology introducing the Fellegi and Holt idea of minimum change. That is, given certain set of matching variables chosen a priori, the methodology look for the closest donor to impute the missing item. The difference between this methodology and the NIM system is the use of a set of matching variables, defined from the start, which are employed to calculate the distance between the recipient and the donor. Additionally, different measure distances are used to find the closest possible donor. This methodology has been developed during the last few years and it is planned to be used for the 2001 Census in the U.K.

SCIA (Automatic Control and Imputation System) is a software developed by the Population and Housing Census and Territory Statistics Service (ISTAT) in Italy (Valente and Massimini, 199?). This methodology involves a mixture of deterministic and probabilistic corrections of persons and housing units records and it also uses the Fellegi and Holt proposal. This methodology was used in the 1991 Italian census, however, new explorations have been carried out in order to create a new system based on the NIM system, and it is also planned to be used in the 2001 Census.

Additionally, the Methodological Department of the Instituto Brasileiro de Geografia e Estadistica in Brasil has implemented a new approach based on regressions trees for imputing income in the population census carried out in the year 2000 (Silva 2001, personal communication).

This approach uses the features provided by the software package SPLUS, for creating the regression trees mentioned, which are essentially binary segmentations based on the complete information as will be explained in successive chapters. After the final partition is created, a random hot deck imputation method is applied independently within each group (terminal nodes of the tree) in order to obtain the final results.

Of special interest is the new development made by National Statistical Office in Korea. The National Statistical Office in Korea is studying the possibility of using CART for creating imputation cells together with a nearest neighbour procedure for imputing missing items with relatively high nonresponse rates in the census (Ryu, J.B. et al 2001).

The main idea of implementing this procedure is to avoid following (2nd and 3rd) call-backs after the census in order to get an answer, and instead, to replace the missing values after the 1st call-back by imputed values.

This procedure has been tested using a pilot survey data for census 2000 for imputing two variables for which nonresponse is higher (2.9% and 2,6%), comparing the results of the 2nd and 3rd call-backs with the imputed values. Analyses show that imputation is strongly recommended after the 1st call-back. Unfortunately, no more information about this study was found available.

Not many offices have published their work done in this area, however the tendency to use donor imputation combined with the minimum change principal proposed by Fellegi and Holt seems to be a common factor in their projects. The preservation of joint distributions is becoming another important factor when doing imputation, which is basically another output of the Fellegi and Holt proposal.

There are many other systems developed by different organisations to improve edit and imputation procedures; however, these are not used for population census data due to their specifications, that is, they were created to solve the missing data problem in specific surveys. Examples of these are: GEIS (Generalised Edit and Imputation System) developed by Statistic Canada; SPEER (Structured Program for Economic Editing and Referrals) developed by U.S. Bureau of the Census; StEPS (Standard Economic Processing System) developed by U.S. Bureau of the Census; Plain Vanilla (General-Purpose Edit and Imputation System for Economic Censuses) developed by U.S. Bureau of the Census; AGGIES (Agriculture Generalised Imputation and Edit System) developed by U.S. Department of Agriculture (Todaro, 1999); Macro View (Graphical Macro Editing System) developed by Statistics Netherlands; CANEDIT developed by Statistic Canada (Bankier, Filliion, Luc & Nadeu, 1994) and DISCRETE developed by U.S. Bureau of the Census.

1.8 LIMITATIONS OF CURRENT METHODS

In general, different imputation methods, some of them used for imputing missing values in census, have different disadvantages. For example, Look up Tables can affect the distribution of the response, the suitability of the methods may be marginal and the use of external data can adversely affect relationship between variables; Mean (or Mode) imputation can distortion relationship between variables and it can modify original distributions and contribute to lowering error estimation; Regression imputation can distortion relationships between variables if they are not included in the model, compresses distributions and lead to problems in estimating valid errors. However, even when these methods can be used for imputing missing information in census data, they are not the most common methods employed in this task.

As mentioned before, the most common method used for imputing census data is a hot deck method, which is normally used in the way of sequential hot deck (Little and Rubin, 1987; Kalton, 1983; Kalton and Kalsbeek, 1992; Madow, et al 1983). This method has the advantage of being very simple and easy to implement, making efficient use of the computing resources as each data file is read only once. Furthermore, when the data is ordered in a way that creates autocorrelation, an additional degree of matching is introduced. However, this method also has some drawbacks that make it a rather inefficient method. One of the problems with the sequential hot deck is the use of very few variables for the classification. This fact does not allow for a very good degree of matching between records, risking the maintenance of relationships between variables. Another important drawback is the possible multiple use of donors, which can contribute to a lowering of the precision of estimates and underestimation of the variances in surveys (census).

Probably one of the most important drawbacks of using hot deck is the aspect related to the preservation of joint distributions. The hot deck method does not necessarily aim for the preservation of joint distributions since the imputation is not made jointly for all the missing values in a record, and even more, the values for filling in the gaps in a specific recipient do not come necessarily from the same donor. This can constitute a very important issue when analysing the data. Sometimes it is not enough to preserve individual marginal distributions, depending on the kind of analysis required.

In fact, one of the most important uses of census data is for examining relationships between variables, for example, how many females for a specific age group who work in a specific area, which makes the maintenance of the joint distributions a very important aspect to take into account when using imputation procedures.

The other imputation method widely used is the Nearest Neighbour. It can be seen from the new imputation development mentioned in the last section that nearest neighbour seems to be a common factor in their projects. This imputation method has some advantages and drawbacks (Lessler and Kalsbeek, 1992; Chen and Shao, 2001; Little and Rubin, 1987). This method identifies the donor that best matches the nonrespondent given a distant function based on a set of auxiliary variables. That makes this method more efficient than other hot deck methods that do not use auxiliary information. It also has the advantage that the data used for imputation is chosen from the same database. However, Nearest Neighbour imputation has also some drawbacks as it does requires considerable computing power since for each recipient the method looks for the closest donor within the dataset. This represents one of the most important drawbacks when imputing census data. It also requires a logical (suitable) choice for measuring "nearness".

1.9 AIMS AND OUTLINE OF THE THESIS

The deficiencies of the current methods described in the previous section, added to the importance of the census data for the statistics in a country, are the main reasons why research about improved methodology for imputing this kind of data has been undertaken here.

The idea is to investigate an alternative method, which uses a different approach to the current available methods, being also simple and efficient.

The method to be investigated in this research involves the use of classification as a first step, followed by imputation within each imputation class. The main idea is to use a classification techniques called CART, which is basically a classification tree technique based on binary segmentation as will be explained in detail in *Chapter* 2, in order to form the imputation classes. After the imputation classes are created, common methods for imputing categorical data are used within each terminal node of the tree. The results of the tree as a whole are compared in order to assess the use of this classification technique in imputation, as well as to compare the different imputation methods used.

The analysis will be carried out for two different targets: the univariate case where a single variable will be imputed, and the multivariate case where two or more single variables will be imputed at the same time by the use of a composite variable. A composite variable is defined by the cross-classification of two or more single variables.

A potential advantage of the proposed approach is the fact that it does not imply the use of complicated procedures or sophisticated technical resources. An aim of the new method is that it should be easy to implement and not require a large amount of time. Moreover, it should not involve high costs.

The use of as many variables as possible (as many as are involved in the relationship) in the classification step is another important aspect of this proposal. The aim is to make the selection of the donor easier and faster.

One of the main aims of the proposed approach is the maintenance of joint distributions, which means upholding correlations between variables when working on the multivariate missing case. The method tries to obtain all the imputations needed for a specific record from the same donor.

Another important aim is for the method proposed to allow for the use of missing covariates in the classification process or even in the imputation process. This is not normally the case. That is, the aim is for the records containing missing information for the x_{ik} variables to be included in the process of growing the tree or to be used as auxiliary information for the imputation.

As this thesis investigates an alternative method for solving the item nonresponse problem in census data, the classification is followed by an imputation procedure. Hence, three different basic imputation methods for categorical data are implemented in this thesis in order to compare the results given the classification. The selection of the methods includes Probability Distribution imputation, Highest Probability (Modal) imputation as well as the use of Nearest Neighbour procedure as it seems to be a common factor in most of the new methodologies created for census data, as mentioned in *Section 1.7*.

This thesis is divided into two main parts, comprising four chapters. Before these two main parts are presented, a description of the tree-based technique with emphasis in CART is given in *Chapter 2*.

The first main part of the thesis is the development of the use of classification trees in the univariate case (when there is only one variable subject to nonresponse). This univariate case consist of two chapters, *Chapter 3* where all the theoretical background and theoretical considerations are explained, and *Chapter 4* which includes the simulation procedures employed and the results obtained for this univariate case.

The second main issue studied in this work is the multivariate case (where more than one variable can be missing at the same time). This case is also divided into two chapters, *Chapter 5* where issues related to the theoretical aspect of the multivariate case are approached, including a description of the different ways in which CART can be used for imputation in the multivariate missing data, and *Chapter 6* which describes the simulation procedure undertaken and the results obtained for this case.

A final chapter, *Chapter 7*, summarises the results and some further work suggested is presented.

CHAPTER 2

TREE-BASED CLASSIFICATION

2.1 INTRODUCTION

The basic principle guiding this thesis is the introduction of a classification method during the imputation process. The procedure to be followed involves classification and imputation, with imputation as the last step within the groups formed by the classification.

Different alternatives that make use of auxiliary information such as Logistic Regression, Linear/Loglinear Regression or Hot Deck within imputation classes can also be employed to impute missing values. In some way, those methods also involve classification as the imputation process control for auxiliary information. However, we refer to classification in this thesis as the use of an auxiliary technique to form imputation classes. Therefore, the approach presented in this thesis consists of two steps, i.e. firstly creating the imputation cells by using a specific classification technique and secondly imputing missing values within those cells using a specific imputation technique.

There are a huge number of methods for classifying elements Hand (1997). Some of them require the estimation of certain parameters (parametric procedures) whilst others do not require those estimations (non-parametric procedures). Some methods require more sophisticated and complex procedures than others. In any case, classification procedures structure the population in a certain way that is useful for researchers in solving specific problems. This structure is constituted by a set of rules based on the values of the variables used for the classification. These variables are measured on a set of units generally called "the learning sample".

The set of rules structuring the population (learning sample) plays two different roles (Hand 1997): one of which is to formulate the class structure, unsupervised classification, and the other operates as a strategy for classifying new elements into those groups already determined, supervised classification.

Given that this thesis deals with the missing information problem, the aim of this research is not only the generation of the classification groups (by the use of the learning sample), which in our problem will be the imputation classes, but also the generation of that set of rules that allows for the classification of units which are not part of the learning sample, which in our case are the records subject to missing information and they are not included in the generation of the original classification.
It has always been assumed that the more homogeneous the population is the better the results of the imputation due to the donor selected. That is, if one can find a procedure that classifies the population in very homogenous sub-groups, the imputation performance should improve in terms of accuracy.

However, the selection of the classification method used is not only based on the accuracy of the classification made. It is also important to say that the use of a procedure that allows the researcher to do the classification without the utilisation of very sophisticated and complex techniques is really convenient when too much information and a very limited time scale for carrying out the task are requested.

Tree-based models have been used over the last several years as an important and useful approach for classifying elements (Gordon, A.D., 1987; Loh, W.Y. and Vanichsetakul, 1988; De Waal, T. 2000; Ryu, J.-B., Kim, Y.-W., Park, J.-W. & Lee, J.-W. 2001). The use of tree-based models release researchers from problems like using complex procedures for parameter estimations or searching for prior information. The power of this methodology in working with large databases and creating accurate and quick classifications as well as some practical factors such as ease of use, recent improvements and computational developments constitute some of the influential factors on the decision for using tree-based models in this analysis.

A tree-based model is a set of classification rules that partitions a data set into mutually exhaustive and non-overlapping subsets (Breiman, L., et al 1984). The rules are defined in terms of the values of a group of explanatory variables. The model is constructed by progressively splitting the data set into smaller subsets that are increasingly more homogeneous with respect to a response variable. This splitting process continues until a stopping criterion is met. Then, the tree-based model is represented by a hierarchical set of splits that eventually lead to the final subsets or "terminal nodes" of the tree.

The Automatic Interaction Detection (AID) program of Sonquist, Baker and Morgan (1971) is one of the first methods for fitting a tree-based model to data. This is based on a recursive

binary algorithm, which successively splits the original data set into two smaller subsets. As mentioned above, these subsets are meant to be more homogeneous subsets than the original one. The partition is made by a succession of sequential binary splits.
A similar recursive binary segmentation algorithm constitutes the bases of the CART (Classification and Regression Tree) program developed by Breiman et al (1984). These ideas have also been implemented in the regression and classification tree analysis modules in S-Plus (Martin and Minardi, 1995). An alternative, non-binary, recursive splitting algorithm underpins the CHAID program (Kass, 1980).

The literature usually refers to two types of tree-based models: Classification Tree models and Regression Tree models. The basic difference between the two models is the scale of measurement of the response variable. In a classification tree model the response variable is assumed to be categorical. In this case, an appropriate measure of homogeneity for categorical data is used in order to determine the splits. In the case of a regression tree, since the response variable is assumed to be continuous, appropriate measures of homogeneity relevant to continuous variables are used to determine the splits in the tree.
In both cases, the explanatory variables can be either categorical or continuos variables.

2.2 CART: THE METHODOLOGY

Classification and Regression Trees (CART) is a segmentation algorithm developed by Breiman et al in the 1980's. This algorithm is known as a binary recursive partition that represents its results in the form of decision trees. It is binary because parent nodes are always split into two subsets (children), and it is recursive because each child could also be treated as a parent and therefore it could also be split. The tree starts with a root node, which is the complete data set of units (universe in our case). This universe is split into two subsets (child nodes) using yes/no questions. Some of the nodes are terminal, which means they are not split any more, while others are not terminal being split until a terminal node is reached.

The main idea behind this classification method is to find decision points for partitioning the universe into mutually exhaustive non-overlapping subsets, given a target or dependent variable (variable for which the classification is done) and given a set of explanatory or independent variables (variables in which the classification is based). These decision points simply represent a set of rules defined in terms of the values of a group of explanatory variables (independent variables), with the model constructed by successively splitting the universe into subsets that are increasingly more homogeneous with respect to a response variable of interest. This splitting process continues until a stopping criterion is met. The tree model is then represented by the hierarchy of splits that eventually lead to the final subsets (terminal nodes) of the tree.

As previously mentioned, the decision points are such the classification is as homogeneous as possible within the terminal nodes created. This means that some categories of the target variable go to one child node and the rest go to the other child node, depending on the values of the independent variable used in order to form groups in which most of the elements come from the same category of the target variable. However, these similarities are not only related to the target variable, but also to the independent variables used in the analysis since the classification is based on their values.

In a classification tree model the response variable is assumed to be categorical, and measures of homogeneity appropriate to categorical data are used to determine the splits in the tree. The independent variables can be either categorical or numerical.

Figure 4.5.3.1 shows an example of a tree that classifies a sample of census records using a variable Primary Activity Last Week as the response and the variables Age, Ethnic Origin and Limiting Long-Term Illness as explanators.

2.3 THEORETICAL FORMULATION

In a more formal way, CART involves specifying the conditional distribution of a dependent variable given a measurement vector $\vec{x}_i$ of independent variables. The binary tree gives a partition of the predictor space in different subgroups for which the distribution of the independent variable is more homogeneous. Each terminal node of the tree corresponds to a region of this partition, and these are determined by splitting rules. At the end, each element of the population is assigned to only one terminal node generating the conditional distribution of the dependent variable at each node.

There are three key elements in CART analysis:

- ✓ splitting each node in a tree
- ✓ deciding when a tree is complete
- ✓ assigning each terminal node to a class outcome (or predicted value for regression)

Each of these elements involves different rules that can be followed in order to obtain the final and optimal tree and they will be explained later.

2.3.1. Definitions

There are some important notation and definitions to be reviewed before starting on a description of CART features.

- ✓ *Response (Dependent variable)* is the variable for which the analysis is being made (variable subject to nonresponse), that is, the variable for which the tree is grown. *Auxiliary variables* (*Independent variables)* are the set of variable used to grow the tree, that is, those variables used as splits.

- ✓ A measurement vector $\vec{x}_i = (x_{i1}, x_{i2}, ..., x_{iK})$ is a vector containing a number of measurements of variables made on a unit i. The collection of all possible measurement vectors in the population U defines the measurement space χ .

- ✓ $C = \{1, ..., J\}$ is the set of classes of the response variable in which each unit may fall into.

- ✓ A classifier is a function $d(\vec{x}_i)$ of $\vec{x}_i$ defined on χ which gives a value between $1, ..., J$ to every measurement vector $\vec{x}_i$.

- ✓ A_j is the subset of χ for which $d(\vec{x}_i) = j$. So, $\chi = \bigcup_j A_j$. Then, a classifier can be defined as a partition of χ into J disjoint subsets $A_1, ...A_j$ for which every element $\vec{x}_i \in A_j$ has j as the predicted class. These disjoint subsets are denominated nodes. These nodes can be terminal (if they are not split anymore) or non-terminal (if they are further split by the process).

- ✓ A Learning Sample is defined by $L = \{(\vec{x}_1, j_1), ..., (\vec{x}_M, j_M)\}$, where $\vec{x}_i$ is a measurement vector with $\vec{x}_i \in \chi$ $j_i \in \{1, ...J\}$, $i = 1, ..., M$, and j_i is the true class for ith unit. M is a subset of the population. In some cases, M can be equal to N (population size).

- ✓ The test sample is a subsample of the learning sample used for estimating the misclassification rate via test sample estimation or cross-validation estimation (see *Section 2.3.6.*). Frequently, this subsample is taken as a 1/3 of the cases. In a 10-fold cross-validation, the total of cases are divided in 10 parts, using a different 1/10 each time as a test sample to estimate the misclassification rate.

- ✓ $C(j_1 \mid j_2)$ is the cost of misclassifying a class j_2 element as a class j_1. $C(j_1 \mid j_2)$ satisfies:

  (a) $C(j_1 \mid j_2) \geq 0, \quad j_1 \neq j_2$, and

  (b) $C(j_1 \mid j_2) = 0, \quad j_1 = j_2$.

- ✓ A split s is defined by a question of the form Is $\vec{x}_i \in A?, \; A \subset \chi$ that sends unit i to the left or right child node depending on the answer of the question.

- ✓ $\pi(j)$ is the set of prior probabilities, that is, the prior probabilities that $y_i = j$, $j = 1,\ldots,J$. These probabilities are either estimated by $\{M_j / M\}$ or pre-specified (i.e. a particular prior distribution for the dependent variable can be specified), with M_j as the number of units in class j in the learning sample. Thus, for a given set of prior $\pi(j)$, $p(j,t) = \pi(j) M_j(t) / M_j$ is taken as the resubstitution estimate for the probability that a unit will both be in class j and fall into node t, where $M_j(t)/M_j$ is the proportion of class j cases in L falling into t.

- ✓ $p(t)$ is the resubstitution estimate of the probability that any case falls into node t, and is defined by $p(t) = \sum_j p(j,t)$, with $p(j,t)$ defined as before. Then, the resubstitution estimate of the probability that a case is in class j given that it falls into node t is given by $p(j \mid t) = p(j,t)/p(t)$ and satisfies $\sum_j p(j \mid t) = 1$.

- ✓ If $\{\pi(j)\} = \{M_j / M\}$, then $p(j,t) = M_j(t)/M(t)$, thus $\{p(j \mid t)\}$ are the relative proportions of class j cases in node t.

- ✓ $\widetilde{T}$ is the set of terminal nodes

- ✓ $|\widetilde{T}|$ is the tree complexity (number of terminal nodes)

- ✓ Cost complexity measure is defined by $R_\alpha(T) = R(T) + \alpha|\widetilde{T}|$, where $R(T)$ is the misclassification rate (see *Section 2.3.6*) and α is the complexity parameter, $\alpha \geq 0$. This cost complexity is then a combination between the misclassification cost of the tree plus a cost penalty for complexity.

- ✓ The impurity function is a function Φ defined on $(p_1,\ldots\ldots,p_J)$, with p_j the proportion of units in class j, $j = 1,\ldots\ldots\ldots,J$, and satisfying

  (a) $p_j \geq 0$ and

  (b) $\sum_j p_j = 1$.

  and it has the following properties:

  (i) Φ is a maximum only at the point $\left(\frac{1}{j},\frac{1}{j},\ldots,\frac{1}{j}\right)$

  (ii) Φ achieves its minimum only at the points $(1,0,\ldots,0),(0,1,\ldots,0),\ldots,(0,0,\ldots,1)$

  (iii) Φ is a symmetric function of $p_1,\ldots p_j$.

2.3.2. Splitting Rules

As mentioned before, CART is known as a binary recursive partitioning. It means that each node is split into two child nodes based on a splitting criterion.
A set S of splits s is generated by a set Q of binary questions in which every value of $\bar{x}_i$ in a node t for which the answer "yes" goes to the descendant left node t_L and every value of $\bar{x}_i$ answering "no" goes to the descendant right node t_R. In general, if the question is {Is $\bar{x}_i \in A?\}$, then $t_L = t \cap A$ and $t_R = t \cap A^c$.

Two different criteria can be found for splitting, Gini criterion and Twoing criterion.

- Gini criterion

This splitting criterion is based on a node impurity measure. The idea is to find the split that reduce the tree impurity defined by $I(T) = \sum_{t \in \widetilde{T}} I(t)$, with $I(t) = \delta\,(t)\,p(t)$, where:

- $\delta(t)$ is a node impurity function defined as $\Phi(p(1 \mid t),\dots\dots,p(J \mid t))$ (relative proportion of class j units in node t),
- $p(j \mid t)$ is defined by $p(j \mid t) = p(j,t)/p(t)$,
- and $p(t)$ is the probability that any case falls into node t where $p(t) = \sum_j p(j,t)$

Another way for defining the impurity function is minimising $I(T)$, which is the same as maximising $\Delta I(s,t) = I(t) - I(t_L) - I(t_R)$ where $\Delta I(s,t)$ represents the decrease in impurity. In other words, to maximise $\Delta\delta(s,t) = \delta(t) - p_L\delta(t_L) - p_R\delta(t_R)$, where p_L is the proportion of units which go from t to t_L and p_R the proportion of units which go from t to t_R. Hence, the best split will be that such as reduce more the misclassification rate $R(t)$.

If $\delta(t)$ is defined as $r(t)$ where $r(t)$ is the misclassification rate for the node t, thus, reducing $I(T)$ could be seen as a reducing the misclassification rate $R(T)$, where $r(t) = \min_i \sum_j c(i \mid j) p(j \mid t)$. Therefore, the best split would be that for which $r(t) - p_L r(t_L) - p_R r(t_R)$ is maximum. This is equivalent to say that $R(t) - R(t_L) - R(t_R)$ is maximum.

There are different criteria for generating the impurity function $\delta(t)$. These are:

Gini Impurity Function

This impurity function has the form $\delta(t) = 1 - SQ$ in which SQ is the sum of squares of the estimated class probabilities $p(j \mid t)$. That is, $\delta(t) = 1 - \sum_j p^2(j \mid t)$.

Given the form of the impurity function, it can be noticed that this function takes values in the interval $[0,1)$. The function reaches the minimum (zero) when the node consists only of a single class, in which case, the node is considered perfectly pure. The function takes the maximum $\left(1 - \frac{1}{J}\right)$ when the node contains equal number of cases for each class.

In this case, it is assumed that all the costs for misclassifying class j_1 as a class j_2 are equal to 1 for all $j_1 \neq j_2$.

Sometimes, the problem requires defining different misclassification costs for different types of misclassification since some of these actions imply more risk than others.

If an unknown element is assigned to a class j with estimated probability $p(j \mid t)$, the expected cost is $\sum_{j_2, j_1} C(j_1 \mid j_2) p(j_1 \mid t) p(j_2 \mid t)$ where $C(j_1 \mid j_2)$ is the misclassification cost. This is the expression used as a Gini node impurity for variable misclassification costs, which is an extension of the original one.

Symmetric Gini Impurity Function

The use of this index assumes symmetry of the misclassification cost matrix. The criterion used is exactly the same used in Gini with the variable cost term. Consequently, the impurity function is $\sum_{j_2, j_1} C(j_1 \mid j_2) p(j_1 \mid t) p(j_2 \mid t)$ as above, where $C(j_1 \mid j_2)$ is a variable misclassification cost but coming from a symmetric matrix.

- *Twoing Criterion*

The second criterion uses a different strategy. This criterion is based on class separation instead of node homogeneity.

Twoing

The basic idea in this method is maximise the difference between the probability that a class j element goes to the left from the probability that the same element goes to the right node.

This criterion defines the towing criterion function as follows:

$$\Omega(s,t) = \frac{p_L p_R}{4} \left[\sum_j \left| p(j \mid t_L) - p(j \mid t_R) \right| \right]^2$$

where, in a two-class criterion for a given split s and a class $C_i(s) = \{ j : p(j \mid t_L) \geq p(j \mid t_R) \}$ that maximises $\Delta\delta\ (s,t,C_1)$, $\max_{C_1} \Delta\delta\ (s,t,C_1) = \Omega(s,t)$.

Then, the best towing split $s^*(C_1^*)$ is given by the split s^* which maximises $\Omega(s,t)$, with C_1^* as $C_1^* = \{ j : p(j \mid t_L^*) \geq p(j \mid t_R^*) \}$. t_L^* and t_R^* are given by the split s^*.

This criterion does not work on the overall impurity measurement of the node maximising $\Delta\delta\ (s,t)$. So, it is not possible to obtain a tree impurity measurement $I(T)$.

Ordered Twoing

Sometimes the categorical variables used are ordered, and it may be desirable to take this characteristic into account. In this case, the twoing criterion can be used but with an additional condition.

The ordered twoing criterion considers a new partition $\{C_1, C_2\}$ of the class $C = \{1,....,J\}$ using the following restriction: $C_1 = \{1,...,j_1\}$, $C_2 = \{j_1 + 1,......,J\}$. It means that there will be a cut-point for which all the classes below this point go to one node and the rest go to the other node. For example, a split can separate classes 1 and 2 from the classes 3 and 4, but cannot separate classes 1 and 3 from the classes 2 and 4.

Therefore, the criterion is given by $\Omega(s,t) = \max\limits_{C_1} \Delta\delta\ (s,t,C_1)$ as for twoing but using the restriction mentioned.

2.3.3. *Class Probability Trees*

This method is used when it is important to estimate the probability that a unit goes to a specific class instead of assigning a class to this element. The probability results are obtained from the within-node distributions of the terminal nodes for the response variable. The tree is always grown using the Gini splitting rule and it is not possible to specify misclassification costs because this tree is not for classifying elements.

The main goal is to estimate the probability distribution of the target variable. It allows specifying prior probabilities. In other words, the basic idea is to estimate $P(j\,|\,\vec{x}_i) = P(y_i = j\,|\,\vec{x}_i = \vec{x}_i\,')$, $j = 1,....,J$, where $\vec{x}_i$ is a measurement vector. This means, to estimate the probability that a case is in class j given an observed vector $\vec{x}_i\,'$ of measurements.

2.3.4. *Class Assignment Rule*

The main objective of the tree is to classify all of the units. This implies the assignment of a class (category) to every unit. This assignment depends on the distribution of the categories of the response variable within each terminal node. Then, once all the elements are allocated to a terminal node, they are assigned a class depending on the node they end up in. Since a single class is allocated to all of the units of each terminal node, these units are treated as they really are from the class assigned. When a tree is finally created, the class assigned to a node identifies all the elements in that particular terminal node.

Thus, each terminal node $t \in \widetilde{T}$ has an assigned class $j \in \{1,.....,J\}$ that is denoted by $j(t)$.

There are two different ways for assigning a class to the units:

- ✓ If the prior probabilities are $\{\pi(j)\} = \{M_j/M\}$, then class assignment is basically the plurality rule for assigning the class. This means t is classified as that class for which $M_j(t)$ is largest.

- ✓ For any other set of priors, $\sum_{j \neq j(t)} p(j\,|\,t)$ is the resubstitution estimate of the probability of misclassification given that a case falls into node t. The class assignment rule $j(t)$ is that rule that minimises this estimate. That is, if $p(j\,|\,t) = \max_i p(i\,|\,t)$ then $j(t) = j$.

2.3.5. Surrogates

One of the important issues related to surrogates is the fact that they allow for the classifying of elements with missing information in the auxiliary variables. This makes possible the use of the whole data set even when missing values are present in the auxiliary variables.
The importance of the surrogate and their uses in the imputation process are explained later in this work.

A surrogate is defined as the alternative split which divides the same set of units in the most similar way to the best split. This similarity is not only related to the number of units in each child node and their internal distributions, but it is also related to which units go to each child node. Surrogates closely mimic the action of the primary split.

In a more formal way, let us take the best split s^* at node t, which divides a set of elements of a node into two different child nodes t_L and t_R. Let us also take any variable x_{ik} with a set of splits S_k and set of complementary splits $\overline{S}_k$. Then, for any split $s_k \in S_k \cup \overline{S}_k$ that divides node t into t_L' and t_R', we have $M_j(LL)$ as the number of units in t sent to left by both s^* and s_k. That is, number of units sent to $t_L \cap t_L'$. Similarly we have $M_j(RR)$ as the number of units in t sent to right by both s^* and s_k. That is, number of units sent to $t_R \cap t_R'$.

The estimated probability that a case falls into $t_L \cap t_L'$ is defined by $p(t_L \cap t_L') = \sum_j \pi(j) \frac{M_j(LL)}{M_j}$. Additionally, the estimated probability that both s^* and s_k sent a case into t to the left, $p_{LL}(s^*, s_k)$ is defined by $p_{LL}(s^*, s_k) = \frac{p(t_L \cap t_L')}{p(t)}$. These probabilities are defined in a similar way for the right node. Additionally, $p(s^*, s_k) = p_{LL}(s^*, s_k) + p_{RR}(s^*, s_k)$.

Thus, a split $\tilde{s}_k \in S_k \cup \overline{S}_k$ is defined as surrogate split on x_{ik} if $p(s^*, \tilde{s}_k) = \max_{s_k} p(s^*, s_k)$ over $S_k \cup \overline{S}_k$. This surrogate split can be interpreted as the split on x_{ik} that predicts in the most accurate way the action of s^*.

Another important aspect related to the use of surrogates is to give a ranking of the auxiliary variables according to their importance in classifying the units. Sometimes, these variables offer trees as accurate as the trees constructed with the original splits.

A measure of the importance of variable x_{ik} is given by $\partial(x_{ik}) = \sum_{t \in T} \Delta I(s_k, t)$, where $\Delta I(s_k, t)$ is the decrease in impurity mentioned in the splitting rules. The quantity used for ranking the importance of the variables is a relative magnitude based on the last equation defined by $100 * \partial(x_{ik}) / \max_k \partial(x_{ik})$, giving a value 100 to the most important variable, and a value between 0 and 100 to the rest.

2.3.6. Estimation of the Misclassification Rate

The misclassification rate is a measure of how accurate the classification is. Given a class structure, the misclassification rate determines the percentage of units misclassified once a class is assigned to each terminal node. These rates could be used for determining, in a way, the predictive power of the tree.

Thus, given a classifier $d(x_{ik})$, its "true misclassification rate" can be denoted by $R^*(d)$, which is defined as the proportion of cases misclassified by $d(x_{ik})$. There are three different ways for estimating the misclassification rate in this method.

First, it is necessary to define a function $\chi(\cdot)$, which is 1 if the condition inside the parentheses is true and 0 otherwise. The ways for calculating the misclassification rate are based on this function.

- ✓ ***Method 1:*** The first way for estimating the misclassification rate, $R(d)$, is called the resubstitution estimate and is defined by $R(d)=\frac{1}{M}\sum_{i}^{M}\chi\left(d\left(\bar{x}_i\right)\neq j_i\right)$. This method uses the data used to construct the classifier, that is, the learning sample L.

- ✓ ***Method 2:*** The second way is called test sample estimation, $R^{ts}(d)$. In this case, L is divided into two groups, L_1 with M_1 elements and L_2 with M_2 elements, $L=L_1\cup L_2$. L_1 (learning sample) is used for constructing the classifiers, and L_2 (test sample) is used for estimating $R^{ts}(d)$, which is defined by $R^{ts}(d)=\frac{1}{M_2}\sum_{(\bar{x}_i,j_i)\in L_2}\chi\left(d\left(\bar{x}_i\right)\neq j_i\right)$. L_1 and L_2 should be considered independent and coming from the same distribution. L_2 is generally 1/3 of L's size.

- ✓ ***Method 3:*** The third way for estimating the misclassification rate is called cross-validation and it is denoted by R^{cv}. In this option, the learning sample L is divided in V subsets of equal size (approximately) denoted by $L_1,\ldots\ldots\ldots\ldots,L_v$. The classifiers $d^{(v)}\left(\bar{x}_i\right)$, $v=1,\ldots\ldots,M$, are constructed with all the elements present in L but not in L_v, ($L-L_v$). An estimation of R^{ts} is given by $R^{ts}\left(d^{(v)}\right)=\frac{1}{M_v}\sum_{(\bar{x}_i,j_i)\in L_v}\chi\left(d^{(v)}\left(\bar{x}_i\right)\neq j_i\right)$ where $M_v\cong M/V$ and none of the elements in L_V are used in the construction of $d^{(v)}$. Then, the final estimation of misclassification rates via cross-validation is $R^{cv}(d)=\frac{1}{V}\sum_{v=1}^{V}R^{ts}\left(d^{(v)}\right)$. All of the V classifiers are constructed using $M(1-1/V)$. In this case, every element is used to construct d and is also used once in the test sample.

The importance of misclassification rates is related to their uses in the generation and pruning processes and other issues about predictive power of the tree.

2.3.7. Pruning Trees

Since CART does not use any procedure to stop growing a tree, a procedure for pruning it is used to obtain the optimal tree. That is, after the maximal tree is grown, the pruning process starts cutting branches based on misclassification rates and a penalty for complex trees.

The purpose in the pruning process is to find the tree from the sequence that minimises $R_{\alpha}(T)$, where $R_{\alpha}(T) = R(T) + \alpha|\tilde{T}|$. In this equation $|\tilde{T}|$ is the complexity of the tree (with $|\tilde{T}|$ as the number of terminal nodes), and α is the complexity parameter. Thus, the last term of the equation can be seen as a cost penalty for complexity, assuming that α is the penalty imposed per additional terminal node.

Then, the process finds the tree that minimises $R_{\alpha}(T)$ for a specified value of α.

The values of α can be either specified by the analyst or automatically obtained by an iterative process carried out by the Software, in which a function of the misclassification rate is minimised (See Breiman, L., et al 1984 for details).

2.3.8. Some Properties

Some of the important attributes that can be mentioned about CART technique are:

- CART does not require the user to make a prior selection of the auxiliary variables. They will be selected from the complete list of variables depending on their power in classifying the units in the population treated.

- Each auxiliary variable can be used in different parts of the tree to detect important interactions between their different combinations. That is, one variable can be used more than once during the growing process.
- CART is invariant with respect to transformations of the auxiliary variables. The use of any transformation will result in the same conclusion.

- Linear combinations of non-categorical auxiliary variables can be used. Also, continuous auxiliary variables can be converted to categorical ones and categorical variables can be collapsed.

- The selection of the variables made by CART can be used for further analyses with linear or logistic regression models, managing a smaller list of variables and prior information about these variables.

- It is nonparametric procedure, which means that does not require any prior specification of the model relating y_i and x_{ik}.

- Missing values for the predictors can be handled by using surrogate splits.
  Surrogates are alternative splits generated when the primary splitting variable is missing. This option permits working with more cases.
  By the use of surrogates, low-cost predictors can be selected. Satisfactory surrogates can generate similar predictions to the original variables. Surrogates can be also useful when the values of some variables are difficult to obtain.

2.4 CART: THE SOFTWARE

CART software was created based on the technique proposed by Breiman, Friedman, Olshen and Stone (1984). This software is a computational version of the original CART methodology, which allows for the rapid growth of trees following all the theoretical processes. It was created in order to simplify the practice of the generation of tree-based models.
There are different versions of this software. The first version was created in 1990. The current version (1999 version, used in this work) introduces new aspects that were not included in previous versions facilitating the use of the original concepts of the technique. In addition, the capacities of data handling have increased from the original version.
This software contains many different options in the growing-tree process; however, not many of them need to be specified during the process since they are specified as default. A brief explanation of some of the most important options follows hereafter:

- The first aspect to be decided is whether a classification or regression tree will be grown. In our application we are only concerned with categorical response variables and so, a classification tree is grown. For growing a classification tree, the labels of the target variable have to be specified.

- The independent variables have to be also defined as categorical or continuous. In the first case, the labels of the categories do not need to be specified but they have to be continuous numbers.

- There are different splitting criteria for growing the tree as explained before. The default method used by CART is the Gini index. However, this could be changed to any of the others above mentioned. The selection of the splitting criterion depends on the kind of variables used for the analysis, both response and auxiliary variables. Depending on the methods used for splitting, the auxiliary variables will be treated as a categorical, ordinal or continuous.

- When continuous variables are used as auxiliary variables, it is possible to use combination of them as splits. In this case, one must specify to the software that a combination of variables is wanted, then, the software decides which linear combinations, if any, are the best splits.

- As explained before, there are also different ways for testing the tree. CART uses Cross-Validation as default considering 10 groups (10-fold cross-validation). It could be changed to any of the other options mentioned in previous sections.

- If any of the variables has any value that should not be included in the analysis for any reason, CART allows the exclusion of specific values from the database used. It is possible to select a subset of values for any of the variables used in the analysis.

- The minimum number of units in each parent node and each terminal node can be specified using an option included in the software; as well as the minimum complexity required (number of terminal nodes required). It is also necessary to specify the maximum number of units used in the learning sample since CART uses 3000 records as default.

- Costs and prior probabilities for the response variable can also be changed. There is a matrix of costs that uses cost 1 for any kind of misclassification as default. This can be changed and symmetrized. In terms of the prior probabilities, equal probabilities are used as default, but this can be changed to the proportion present in the learning sample, test sample, the data, a mixture of them or another specified distribution.
- In terms of surrogates, it is possible to select how many surrogates are wanted to appear and to decide whether or not all of them have the same weight for the variable importance.

- Sometimes, the data set used is too large for using any determined version of the software. In these cases, there are a number of options to solve this problem. Test sample sizes and learning sample sizes can be modified. The depth of the tree, the number and size of the nodes can also be changed. Also, a subsampling of the data set can be used for the analysis.

CHAPTER 3

UNIVARIATE CASE THEORETICAL FRAMEWORK

3.1. INTRODUCTION

This chapter contains the theoretical formulation of the univariate work undertaken in this thesis. Here, the univariate case is explained including modelling description, the use of classification trees, imputation methods used and estimation of population quantities. Additionally, properties of the estimators are studied.

It is important to specify that in this work the terms univariate and multivariate refer to the number of variables subject to nonresponse, no matter how many variables are fully observed. This is, the univariate case refers to the situation where only one variable is subject to nonresponse and the multivariate case refers to the situation where more than one variable are subject to nonresponse. The theoretical formulation for the multivariate case is presented in *Chapter 5*.

3.2 NOTATION

Using the notation defined in *Chapter 1*, let U be a finite population of N elements $U=\{U_i; i=1,2,...,N\}$. Let $\mathbf{Y}=(y_i)$ be a $(Nx1)-vector$ of response variable, where y_i represents the $i\ th$ element and let $\mathbf{X}=(x_{ik})$ be a $(NxK)-matrix$ of auxiliary variables where x_{ik} represents the $k\ th$ variable for $i\ th$ the element.

In this case, $\mathbf{Y}$ can be represented as a vector of N values y_i, $\mathbf{Y} = (y_1, y_2, ..., y_n, ..., y_N)^t$; $\mathbf{X}$ can be represented as $\mathbf{X} = (\vec{X}_1, \vec{X}_2, ..., \vec{X}_k, ..., \vec{X}_K)$, where $\vec{X}_k = (x_{1k}, x_{2k}, ..., x_{Nk})^t$ is a vector of N values x_{ik}.

Assuming that $\mathbf{Y}$ is subject to nonresponse and $\mathbf{X}$ is fully observed, we have $\mathbf{R} = (r_i)$ as $(N\text{x}1) - vector$ of indicator variables for $\mathbf{Y}$ identifying whether or not y_i is missing. That is, $r_i = \begin{cases} 1 & if\ y_i\ is\ observed \\ 0 & otherwise \end{cases}$. $\mathbf{R}$ can be represented as a vector of N values r_i, $\mathbf{R} = (r_1, r_2, ..., r_n, ..., r_N)^t$.

It is also assumed the population is fully enumerated (no sample is taken).
The data take the form

$$\begin{array}{cccccc}
\vec{X}_1 & \vec{X}_2 & \dots & \vec{X}_K & \mathbf{Y} & \mathbf{R} \\
 & & & & & \\
x_{11} & x_{12} & \dots & x_{1K} & y_1 & 1 \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
x_{m1} & x_{m2} & \dots & x_{mK} & y_m & 1 \\
x_{m+1,1} & x_{m+1,2} & \dots & x_{m+1,K} & 0 & 0 \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
x_{N1} & x_{N2} & \dots & x_{NK} & 0 & 0
\end{array}$$

where the zeros represent the missing values in the population and x_{ik} and y_i are specific values for a specific realisation of the model, with m the number of records for which $\mathbf{Y}$ is observed (measured), the zeros represent the missing values and N is the number of elements in the population. That is, we take without loss of generality $r_1 = ... = r_m = 1$ and $r_{m+1} = ... = r_N = 0$.

3.3 MODEL DESCRIPTION

Under the model-based approach, x_{ik} and y_i are considered random variables with distribution $f(\mathbf{X},\mathbf{Y}\,|\,\boldsymbol{\theta})$ indexed by the parameter (sets of parameters) $\boldsymbol{\theta}$.

As the response process can be seen as a random process, the response outcome $\mathbf{R}$ is also included as a random variable with distribution $f(\mathbf{R}\,|\,\mathbf{X},\mathbf{Y},\boldsymbol{\varphi})$.
Providing that $\mathbf{X}$ is fully observed and $\mathbf{Y}$ is subject to nonresponse, the full form of the distribution can be written as $f(\mathbf{X},\mathbf{Y},\mathbf{R}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})$ indexed by the parameters (sets of parameters) $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$, with $\mathbf{R}$ as a response indicator.

The joint distribution of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{R}$, $f(\mathbf{X},\mathbf{Y},\mathbf{R}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})$, can be decomposed as the product of the probability distribution of $\mathbf{X}$ and $\mathbf{Y}$ indexed by the parameter (set of parameters) $\boldsymbol{\theta}$ and the conditional distribution of $\mathbf{R}$ given $\mathbf{X}$ and $\mathbf{Y}$ (the distribution for the missing data mechanism) indexed by the parameter (set of parameters) $\boldsymbol{\varphi}$. That is,

$$f(\mathbf{X},\mathbf{Y},\mathbf{R}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}) \quad = \quad f(\mathbf{X},\mathbf{Y}\,|\,\boldsymbol{\theta}) \quad f(\mathbf{R}\,|\,\mathbf{X},\mathbf{Y},\boldsymbol{\varphi}) \qquad (1)$$

Since $\mathbf{Y}$ is subject to nonresponse, we can write $\mathbf{Y}=(\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}})$, where $\mathbf{Y}_{\mathbf{obs}}$, which is a vector of size $mx1$, represents the observed values of $\mathbf{Y}$ and $\mathbf{Y}_{\mathbf{mis}}$, which is a vector of size $(N-m)x1$, represents the missing values of $\mathbf{Y}$.

Therefore, the distribution $f(\mathbf{X},\mathbf{Y},\mathbf{R}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})$ can be written as $f(\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}},\mathbf{R}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})$.
Furthermore, equation (1) can be written as

$$f(\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}},\mathbf{R}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})=f(\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}}\,|\,\boldsymbol{\theta})f(\mathbf{R}\,|\,\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}},\boldsymbol{\varphi}).$$

The distribution of the observed data can be obtained by integrating $\mathbf{Y}_{\mathbf{mis}}$ out of the joint distribution of $\mathbf{X}$, $\mathbf{Y}\,and\,\mathbf{R}$. That is, $f(\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{R})=\int f(\mathbf{X},\mathbf{Y},\mathbf{R})\,d\mathbf{Y}_{\mathbf{mis}}$. More specifically, $f(\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{R}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})=\int f(\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}}\,|\,\boldsymbol{\theta})\,f(\mathbf{R}\,|\,\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}},\boldsymbol{\varphi})\,d\mathbf{Y}_{\mathbf{mis}}$.

Assumptions about the model are normally made in order to obtain valid estimation. One of the most common assumptions is that the missing values are "*missing at random*", MAR (Little and Rubin 1987).

As explained in *Chapter 1*, the data is said to be missing at random if the response indicator $\mathbf{R}$ does not depend on the missing values of $\mathbf{Y}$, $\mathbf{Y}_{\text{mis}}$. That is, MAR holds if $f(\mathbf{R} \mid \mathbf{X}, \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \varphi) = f(\mathbf{R} \mid \mathbf{X}, \mathbf{Y}_{\text{obs}}, \varphi)$.

Then, assuming that MAR holds, and given that the actual observed data is $(\mathbf{X}, \mathbf{Y}_{\text{obs}}, \mathbf{R})$, we now have

$$f(\mathbf{X}, \mathbf{Y}_{\text{obs}}, \mathbf{R} \mid \boldsymbol{\theta}, \varphi) = f(\mathbf{X}, \mathbf{Y}_{\text{obs}} \mid \boldsymbol{\theta}) f(\mathbf{R} \mid \mathbf{X}, \mathbf{Y}_{\text{obs}}, \varphi).$$

In this case, the common maximum likelihood procedure used for fully observed data can be used for estimating the parameter $\boldsymbol{\theta}$ required when the data is incomplete (data with missing values). That is, $\boldsymbol{\theta}$ can be estimated by maximising $f(\mathbf{X}, \mathbf{Y}_{\text{obs}} \mid \boldsymbol{\theta})$ the observed data. Hence, the missing data mechanism is ignorable, that is the second part of the right hand side of the last equation can be ignored in the estimation of $\boldsymbol{\theta}$.

Thus, if MAR holds, inferences for $\boldsymbol{\theta}$ are based on the likelihood function $L(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Y}_{\text{obs}})$, which is a function of $\boldsymbol{\theta}$ proportional to $f(\mathbf{X}, \mathbf{Y}_{\text{obs}} \mid \boldsymbol{\theta})$.

3.4 USING CLASSIFICATION TREES

One of the important issues underlying this work is the use of classification as a first step for imputation. Here, the classification is employed to form the classes in which the imputation will be carried out.

As described in *Chapter 2*, the classification method used is called CART (Classification and Regression Trees) and consists of grouping records depending on a set of values of the variables x_{ik}, independent or explanatory variables, which are assumed to be fully observed in this chapter. These groups are called terminal nodes or classification groups and they are used as imputation classes. Additionally, these terminal nodes are expected to be exclusive and exhaustive groups.

In practice, the explanatory variables can also be subject to nonresponse. However, let us assume fully observed explanatory variables in this section.

As in *Chapter 2*, let t represent the node, which is defined by a set of values of the explanatory variables identifying the classification groups. In this case we will use t as a subscript for representing only terminal nodes, with $t \in \{1,2,...,t,...,T\}$, and T equal to the total number of terminal nodes for a specific tree.

As also defined in *Chapter 2,* a measurement vector $\vec{x}_i = (x_{i1}, x_{i2}, ..., x_{iK})$ is a vector containing a number of measurements made on unit i, where $\mathbf{X} = (x_{ik})$ is the matrix containing the values of these variables fully observed defined at the beginning of the chapter. The collection of all possible measurement vectors defines the measurement space χ, with $\chi = \{\vec{x}_i\,; i = 1,2,...,N\}$. We define χ_t as the set of measurement vectors belonging to a specific terminal node with $\chi = \chi_1 \cup \chi_2 \cup ... \cup \chi_T$.

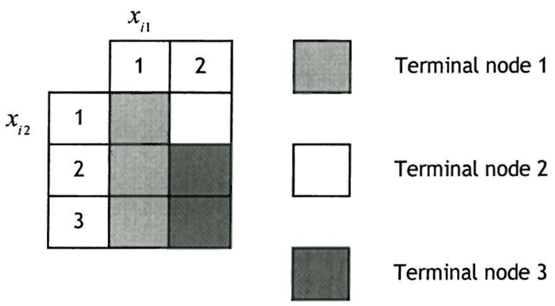
Under the model-based assumption, we also write the probability function of $\mathbf{Y}$, which is subject to nonresponse, given the terminal node defined by χ_t as $f(\mathbf{Y} \mid \vec{x}_i \in \chi_t)$. That is, the probability function of $\mathbf{Y}$ given a set of values of the explanatory variables identifying that terminal node (classification group). To simplify the notation we write $f(\mathbf{Y} \mid \vec{x}_i \in \chi_t) = f(\mathbf{Y} \mid t)$.

Since all the variables used in this work are categorical, we write $f_{\mathrm{Y}}(j \mid t) = P(y_i = j \mid \vec{x}_i \in \chi_t)$ as the probability that y_i takes the value j in a terminal node t. Here, $j = \{1,...,J\}$ represent different categories of the variable y_i.

Example

To illustrate this, let y_i be the dependent variables taking values 0 or 1, and $\mathbf{X} = \{\vec{X}_1, \vec{X}_2\}$ two independent vector of variables with x_{i1} taking values 1 or 2 and x_{i2} taking values 1, 2 or 3. Then, the measurement space χ, is defined by $\chi = \{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3)\}$.

Suppose that the classification tree divides the group of elements in three terminal nodes, with, the space χ consisting of $\chi = \chi_1 \cup \chi_2 \cup \chi_3$, defined by $\chi_1 = \{\vec{x}_i\,; x_{i1} = 1\}$; $\chi_2 = \{\vec{x}_i\,; x_{i1} = 2 \; and \; x_{i2} = 1\}$ and $\chi_3 = \{\vec{x}_i\,; x_{i1} = 2 \; and \; (x_{i2} = 2 \; or \; x_{i2} = 3)\}$. As it can be seen in the following picture, these groups are exclusive and exhaustive groups.



Providing the set of classification rules defining the three terminal nodes, these terminal nodes can be written as $\chi_1=\{(1,1),(1,2),(1,3)\}$, $\chi_2=\{(2,1)\}$, and $\chi_3=\{(2,2),(2,3)\}$. Then, the probability that y_i takes value j given the terminal node χ_t, or equivalent t, can be written as $f_{\mathbf{Y}}(j\,|\,t)=\Pr(y_i=j\,|\,t)=P^{\xi}_{jt}$ with $j=0,1$, $t=1,2,3$ and

$$P^{\xi}_{jt}=\frac{P(y_i=j,\vec{x}_i\in\chi_t)}{P(\vec{x}_i\in\chi_t)}.$$

The inclusion of the classification groups introduces a new factor to the distributions mentioned so far. That is, given a specific classification, let us write $f(\mathbf{Y}\,|\,\chi_t,\boldsymbol{\theta})=f(\mathbf{Y}\,|\,t,\boldsymbol{\theta})$ as the probability function of $\mathbf{Y}$ given the terminal node t indexed by set of parameters $\boldsymbol{\theta}=(\theta_1,\theta_2,\ldots,\theta_T)$.

Now, the joint distribution of $\mathbf{Y}$ and $\mathbf{R}$ given a terminal node t can be written as $f(\mathbf{Y},\mathbf{R}\,|\,t,\boldsymbol{\theta},\boldsymbol{\varphi})$ and as in equation (1), this can be decomposed as $f(\mathbf{Y},\mathbf{R}\,|\,t,\boldsymbol{\theta},\boldsymbol{\varphi})=f(\mathbf{Y}\,|\,t,\boldsymbol{\theta})f(\mathbf{R}\,|\,\mathbf{Y},t,\boldsymbol{\varphi})$.

As before, if MAR holds and assuming independence between units, then $f(y_i\,|\,r_i=0,t,\boldsymbol{\theta},\boldsymbol{\varphi})=f(y_i\,|\,r_i=1,t,\boldsymbol{\theta},\boldsymbol{\varphi})=f(y_i\,|\,t,\boldsymbol{\theta})$ (Little and Rubin, 1987).

We have been holding the assumption of MAR given the x_{ik} variables. However, since we are using classification trees for forming the imputation classes, that is, the imputation is made within terminal nodes, we now want to assume MAR within terminal nodes.

We define the distribution of $\mathbf{Y}$ given $\mathbf{X}$ as $f_{\mathbf{Y}}(j\,|\,x)$ with $f_{\mathbf{Y}}(j\,|\,x)=\Pr(y_i=j\,|\,\mathbf{X}=x)$, $x\in\chi$; and the distribution of $\mathbf{X}$ as $f_{\mathbf{X}}(x)$ with $f_{\mathbf{X}}(x)=\Pr(\mathbf{X}=x)$, $x\in\chi$.

Then, by definition of terminal nodes we have $\Pr(y_i = j \mid t) = \Pr(y_i = j \mid x \in \chi_t)$ where

$$\Pr(y_i = j \mid x \in \chi_t) = \frac{\sum_{x \in \chi_t} \Pr(y_i = j, X = x)}{\sum_{x \in \chi_t} \Pr(X = x)} = \frac{\sum_{x \in \chi_t} f_{\mathrm{Y}}(j \mid x) f_{\mathrm{X}}(x)}{\sum_{x \in \chi_t} f_{\mathrm{X}}(x)}$$

If $\mathbf{Y}$ depends upon $\mathbf{X}$ only via the terminal node so that $f_{\mathrm{Y}}(j \mid x) = f_{\mathrm{Y}}(j \mid t)$ for all $x \in \chi_t$, then, $\Pr(y_i = j \mid x \in \chi_t)$ $= \dfrac{\sum_{x \in \chi_t} f_{\mathrm{Y}}(j \mid t) f_{\mathrm{X}}(x)}{\sum_{x \in \chi_t} f_{\mathrm{X}}(x)}$ $= f_{\mathrm{Y}}(j \mid t)$. That is, $\Pr(y_i = j \mid x \in \chi_t) = f_{\mathrm{Y}}(j \mid t)$, which implies assuming MAR within terminal nodes. That is what we will assume from now on.

3.5 IMPUTATION METHODS

Once the classification tree is constructed, each imputation method is applied at each terminal node. Three different imputation methods are considered. These methods are common for categorical data, which is the kind of data used in the analysis. A description of those methods follows.

Before describing the imputation methods used, let us remember some important concepts defined in *Chapter 1* that are also used in this section.

The independent (explanatory) variables are those used for the classification (variables x_{ik}). They are normally fully observed, while the dependent (response) vector of variables is that vector for which the classification is done ($\mathbf{Y}$) and it is subject to nonresponse.

Additionally, *recipients* are those records containing missing values (records to be imputed), while *donors* are those records which information is completely observed (records from which values to be imputed are taken) and used to impute missing values.

Only one vector of variable is subject to nonresponse in the univariate case while many vectors of variables can be used as explanatory variables. As before, the explanatory variables x_{ik} are considered fully observed in this section. Different approaches for imputing missing values using donors with missing items in some of the exploratory variables will be explained later in this work.

Given the response indicator $\mathbf{R}$, we define r as the set of observed elements, $r=\{i;r_i=1\}$ with $r_t=\{i;r_i=1\,|\,\bar{x}_i\in\chi_t\}$ and $r=r^1\cup\ldots\cup r^T$. The same for the unobserved elements, $nr=\{i;r_i=0\}$ with $nr_t=\{i;r_i=0\,|\,\bar{x}_i\in\chi_t\}$ and $nr=nr^1\cup\ldots\cup nr^T$.

Additionally, we define N_t as the number of records in terminal node t and m_t as the number of observed records in terminal node t, with $m=\sum_{t=1}^{T}m_t$ total of observed elements in the population and $N=\sum_{t=1}^{T}N_t$ total of elements in the population.

1. Probability Distribution Method

As in any case with missing values, we want to impute the missing value of y_i when this is missing from $f(y_i\,|\,r_i=0,t,\theta)$.

Assuming that MAR holds and assuming independence between the units, we can write $f\left(\mathbf{Y}_{\mathbf{mis}}\,|\,\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{R}\right)=f\left(\mathbf{Y}_{\mathbf{mis}}\,|\,\mathbf{X},\mathbf{Y}_{\mathbf{obs}}\right)=f\left(\mathbf{Y}_{\mathbf{mis}}\,|\,\mathbf{X}\right)$, so that

$f\left(y_i\,|\,r_i=0,\bar{x}_i\right)=f\left(y_i\,|\,r_i=1,\bar{x}_i\right)$.

For a tree model it is supposed that $f(y_i\,|\,r_i=0,t,\boldsymbol{\theta})=f(y_i\,|\,r_i=1,t,\boldsymbol{\theta})$ where $f(y_i\,|\,r_i=1,t,\boldsymbol{\theta})$ is the probability distribution of the observed values given the terminal node t, $f(y_i\,|\,r_i=0,t,\boldsymbol{\theta})$ is the probability distribution of the missing values given the terminal node t. Then, the probability distribution method works as follows: given a specific tree, and for each terminal node, the probability distribution of the observed values of the response variable $f(y_i\,|\,r_i=1,t,\boldsymbol{\theta})$ determines the values to be imputed.

Since variables y_i are categorical, we write $\Pr(y_i=j\,|\,r_i=1,t,\boldsymbol{\theta})=P_{jt}^{\xi}$, where j represents the categories of the variable y_i with $j=\{1,2,\ldots,J\}$. We estimate P_{jt}^{ξ} by the observed proportion of cases with category j for the variable y_i in terminal node t in the population, $\hat{p}_{jt}$.

In summary, the probability distribution of $\mathbf{Y}$ for the missing data is assumed to be equal to the probability distribution of $\mathbf{Y}$ for the observed data, which is estimated by $\hat{p}_{jt}$.

Example

To illustrate this method, suppose that a particular terminal node t of the tree used has the following observed distribution for the variable to be imputed

$$\hat{p}_{jt} = \begin{cases} .108; & j=1 \\ .596; & j=2 \\ .242; & j=3 \\ .054; & j=4 \end{cases}$$

This is, for that specific terminal node, 10.8% of the records of the observed variable have category 1, 59.6% of the records have category 2, 24.2% of the records have category 3 and only 5.4% of the records have category 4.

Then, records with missing values for the variable to be imputed that end up in that terminal node will be imputed with category $j=1$ with probability 0.108, with category $j=2$ with probability 0.596, and so on.

2. Highest Probability Method (or Modal Imputation)

Under the same assumptions made for the probability distribution method, that is, MAR holds and independence between units, and given a specific tree, this method imputes the value that is "most likely" in that specific terminal node (i.e. has the highest probability) to all of the records with missing values. Thus, the value to be imputed will be j^*, satisfying $\hat{p}_{j^*t} \geq \hat{p}_{jt}$, for all categories j of the response variable.

Then, in this case, the imputation takes the value $\hat{y}_i = j^*$

It could be more than one j^* value satisfying this condition. In this case, the method selects one of the categories randomly with equal probabilities.

Example

An illustration of this method can be given by the following example. Suppose that the results for the variable to be imputed at one specific terminal node t of the tree has the same distribution as in the last example, that is,

$$\hat{p}_{jt} = \begin{cases} .108; & j=1 \\ .596; & j=2 \\ .242; & j=3 \\ .054; & j=4 \end{cases}$$

Then, all the records with missing values for $\mathbf{Y}$ that end up in that specific terminal node will be imputed as category 2, given that this category has the highest probability in that node.

3. Nearest Neighbour Method

Given a specific tree and for each terminal node individually, distances between the recipient and each possible donor are calculated and the "nearest" donor defines the imputed value for that particular recipient. The nearest donor is determined by the set of independent variables. That is, the distance between the two records (recipient and possible donor) is calculated by adding one to the distance function every time different values are found between them for the independent variables.

Then, given a recipient i' with values $x_{i'k}$, $k=1,2,\ldots,K$ for the vector $\vec{X}_{i'}$, a donor i with value y_i for $\mathbf{Y}$ and values x_{ik}, $k=1,2,\ldots,K$ for the vector $\vec{X}_i$ is that record which satisfies $\min_i[d_{i'i}]$ with $i' \in nr^t$ and $i \in r^t$, and $d_{i'i} = \sum_{k=1}^{K} I(x_{i'k} \neq x_{ik})$

Then, the missing value $y_{i'}$ will be imputed with the observed value y_i from the donor i, $\hat{y}_{i'} = y_i$.

In this case we define A_i as the number of times unit i is used as donor, therefore, $A_i = \sum_{i \in nr_t} I(d_{i'i} \leq d_{i'l} \; for\ all\ l \in r_t)$.

It is important to point out that in this case a record can be used more than once as a donor. This means that if a specific record has the least distance to two different recipients, this record could be used as a donor to fill in the missing values for both of the recipients.
Moreover, when a recipient has the same distance to two different donors, one of the donors is randomly selected with equal probabilities.

Example

An example of this method can be given by the following situation. Suppose that there are six units in a specific terminal node, one of them have variable y_i missing. Each unit contains five values for the five different independent variables as follows:

Units	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}
U1		0	6	2	0	8
U2	1	1	1	2	1	9
U3	4	1	9	3	1	4
U4	5	0	6	3	1	8
U5	2	1	5	2	0	7
U6	2	0	4	3	1	6

In this case, the distances are calculated between unit 1 (recipient) and the rest of the units (possible donors).

In the first comparison, it can be seen that the difference between units 1 and 2 is equal to four. This is because four of the five variables x_{ik} used for the comparison have different values from units 1 to 2, i.e. the recipient and the first possible donor have different values between variables x_{i1}, x_{i2}, x_{i4} and x_{i5} and equal values for variable x_{i3}. The same can be applied to the rest of the possible donors.

A table containing the distances between the recipient and all the possible donors is hereafter

Comparison	U1 - U2	U1 - U3	U1 - U4	U1 - U5	U1 - U6
Distance value	4	5	2	3	4

Given the distances shown in the last table, we can say that the donor used for imputing the recipient will be the unit number 4 (U4) since it produces the smallest distance to the recipient. It means, $\left[\sum_{k=1}^{5} I(x_{1k} \neq x_{ik})\right]$ $i = 2,\ldots,6$ is minimum for U4. Then, since the value for the y_i variable in U4 is five, the imputation value will be 5.

3.6 ESTIMATION OF POPULATION QUANTITIES

Inferences can be done for finite population quantities or for superpopulation parameters depending on the kind of analysis required.

It is more commonly the case for census data, that researchers are interested in making inferences about the quantities that characterise the population, rather than the superpopulation underlying that population. Both cases are important, however. Since the aim of this work is more a descriptive analysis than an analytic analysis, the estimation will be concentrated on population quantities.

Given a finite population $U = \{U_i; i = 1, 2, ..., N\}$ of N elements and vector $\mathbf{Y}$ of interest, taking values $y_i\ ; i = 1, 2, ..., N$, the aim is to estimate a population quantity, for example, the total $Y = \sum y_i$.

A population quantity can be represented as a function of the population values, for example, $g(y_1, ..., y_N)$.

Suppose that the quantity of interest is the number of cases i with category j for $\mathbf{Y}$, which in our case is a categorical variable taking values j from $j = \{1, ..., J\}$. Here, the number of cases in category j of $\mathbf{Y}$ can be written as $g_j = \sum_U I(y_i = j)$.

Since not all the data is observed, this population quantity can be estimated as follows

$$\hat{g}_j = \sum_r I(y_i = j) + \sum_{nr} I(\hat{y}_i = j)$$

which is the same as $\hat{g}_j = \sum_{i=1}^{m} I(y_i = j) + \sum_{i=m+1}^{N} I(\hat{y}_i = j)$, where y_i is the value of $\mathbf{Y}$ for the unit i if present and $\hat{y}_i$ is the imputed value of $\mathbf{Y}$ for the unit i if missing.

The first part of the expression, $\sum_r I(y_i = j)$, can be calculated from the observed data, while the second part of the expression, $\sum_{nr} I(\hat{y}_i = j)$, depends on the imputation method used, since the imputed values $\hat{y}_i$ are determined by the imputation method used.

3.7 PROPERTIES OF THE ESTIMATORS

Let us examine the properties of the estimator of the total defined in the last section. In this work, we refer to bias and variance as properties of the estimator.

In order to examine these properties, let us assume the following statements:

$f(Y \mid \vec{\mathbf{X}})$ represents the probability distribution of the Y given $\vec{\mathbf{X}}$, where $\vec{\mathbf{X}}$ is fully observed. Both, Y and X_k are categorical variables.

We are assuming model-based approach, where $P^{\xi}(y_i = j \mid \chi_t)$ represents the probability that y_i takes the value j in a specific terminal node t given by the model ξ .

We are also assuming that $\mathbf{Y}$ is missing at random within terminal nodes, that means, $\Pr(y_i = j \mid t, r_i = 1) = \Pr(y_i = j \mid t, r_i = 0)$ as defined in *Section 3.4*, and holding r_i and x_i fixed.

Notation

Let us summarise some notation used in the assessment of expectation and variance

$\hat{y}_i$ imputation or imputed value

$j = \{1,...,J\}$ categories of the Y variable

$t = \{1,...,T\}$ terminal nodes

r respondents, $r = \{1,...,m\}$

r_t respondents in node t, $r_t = \{1,...,m_t\}$

nr nonrespondents, $nr = \{m+1,...,N\}$

nr_t nonrespondents in node t, $nr_t = \{m_t + 1,...,N_t\}$

$m = \sum_{t=1}^{T} m_t$ total of observed elements in the population

$N = \sum_{t=1}^{T} N_t$ total of elements in the population

ξ represents the model under the assumption mentioned above

P_{jt}^{ξ} probability that variable Y takes the value j in terminal node t given the model, that is, $P^{\xi}(y_i = j \mid t)$, or equivalently $P^{\xi}(y_i = j \mid \chi_t)$. This probability is assumed the same for all y_i within a terminal node t.

$\hat{p}_{jt}$ proportion of cases in category j in node t using the observed data

E_{ξ} expectation with respect to the model

E_I expectation with respect to the imputation process

As said at the beginning of the section, we refer to bias and variance as properties of the estimators.

Since the estimator of interest is the total of units i with category j for a specific categorical variable, which we define as $g_j = \sum_U I(y_i = j)$, we want to examine the bias and variance of the different between the real total and its estimator. That is, we will be looking at the expected value and variance of that difference, $E_{\xi I}\left[\hat{g}_j - g_j\right]$ and $V(\hat{g}_j - g_j)$. Additionally, we will search for an estimator of the variance obtained for each method and determine if this is unbiased by looking at the difference between $V_\xi E_I(\hat{g}_j - g_j) + E_\xi V_I(\hat{g}_j - g_j)$ and the actual variance.

In our case, we can have two different sources of random variation, one is the model and the other is a stochastic random variation coming from some imputation methods. Therefore, we take the expected value not only with respect to the model, but also with respect to the imputation process. Later in this chapter, we will examine a model free approach in which the random variation will come from the response mechanism and the imputation methods. In this latest case, no assumptions about the response mechanism are made as in the case presented hereafter.

3.7.1. Probability Distribution Imputation Case

Let us define the estimator of the total of $\mathbf{Y}$ when using probability distribution method for imputing. As explained in *Section 3.5*, this method sets $\hat{y}_i = j$ with probability p_{jt} estimated by $\hat{p}_{jt}$. Then, the estimator of the total for the $\mathbf{Y}$ can be written as follows

$$\hat{g}_j = \sum_r I(y_i = j) + \sum_{nr} I(\hat{y}_i = j) = \sum_t \sum_{r_t} I(y_i = j) + \sum_t \sum_{nr_t} I(\hat{y}_i = j)$$

which can be approximated by using the expression

$$\sum_t \sum_{r_t} I(y_i = j) + \sum_t \hat{p}_{jt}(N_t - m_t)$$

since $\hat{p}_{jt}$ is the observed proportion of cases for category j of $\mathbf{Y}$ in terminal node t in the population, that is, $\hat{p}_{jt} = \dfrac{\sum_{r_t} I(y_i = j)}{m_t}$, with m_t equal to the number of observed cases for $\mathbf{Y}$ in terminal node t.

That is, $\hat{y}_i = j$ with estimated probability $\hat{p}_{jt}$ if $i \in t$ and y_i is missing.

Bias with respect to the model

To assess the bias of $\hat{g}_j$ as an estimator of g_j let us calculate the expected value of $\hat{g}_j - g_j$.

$$E_{\xi I}\left[\hat{g}_j - g_j\right] = E_{\xi I}\left[\sum_t\sum_{i=1}^{m_t} I(y_i=j) + \sum_t \sum_{i=m_t+1}^{N_t} I(\hat{y}_i=j) - \sum_t\sum_{i=1}^{N_t} I(y_i=j)\right]$$

$$= E_\xi\left[\sum_t\sum_{i=1}^{m_t} I(y_i=j)+\sum_t\sum_{i=m_t+1}^{N_t} E_I\left[I(\hat{y}_i=j)\right] - \sum_t\sum_{i=1}^{N_t} I(y_i=j)\right]$$

$$= E_\xi\left[\sum_t\sum_{i=1}^{m_t} I(y_i=j)+\sum_t\sum_{i=m_t+1}^{N_t} \hat{p}_{jt} - \sum_t\sum_{i=1}^{N_t} I(y_i=j)\right]$$

$$= E_\xi\left[\sum_t\sum_{i=1}^{m_t} I(y_i=j)+\sum_t (N_t-m_t)\sum_{i=1}^{m_t}\frac{I(y_i=j)}{m_t} - \sum_t\sum_{i=1}^{N_t} I(y_i=j)\right]$$

$$= E_\xi\left[\sum_t\left[\sum_{i=1}^{m_t} I(y_i=j)+(N_t-m_t)\sum_{i=1}^{m_t}\frac{I(y_i=j)}{m_t}-\sum_{i=1}^{N_t}I(y_i=j)\right]\right]$$

$$= \sum_t\left[E_\xi\left(\sum_{i=1}^{m_t}I(y_i=j)\right)+E_\xi\left((N_t-m_t)\sum_{i=1}^{m_t}\frac{I(y_i=j)}{m_t}\right)-E_\xi\left(\sum_{i=1}^{N_t}I(y_i=j)\right)\right]$$

$$= \sum_t\left[\left(\sum_{i=1}^{m_t}E_\xi\left(I(y_i=j)\right)\right)+\left((N_t-m_t)\left(\sum_{i=1}^{m_t}\frac{E_\xi\left(I(y_i=j)\right)}{m_t}\right)\right)-\left(\sum_{i=1}^{N_t}E_\xi\left(I(y_i=j)\right)\right)\right]$$

$$= \sum_t\left[\left(m_tP_{jt}^{\xi}\right)+\left((N_t-m_t)m_t\frac{P_{jt}^{\xi}}{m_t}\right)-\left(N_tP_{jt}^{\xi}\right)\right] = \sum_t\left[m_tP_{jt}^{\xi}+N_tP_{jt}^{\xi}-m_tP_{jt}^{\xi}-N_tP_{jt}^{\xi}\right] = 0$$

Hence, $\hat{g}_j$ is an unbiased estimator of g_j under the model assumptions.

Variance with respect to the model

The variance of the difference between the estimator of g_j and the parameter can be expressed as follows

$$V(\hat{g}_j - g_j) = V_\xi E_I(\hat{g}_j - g_j) + E_\xi V_I(\hat{g}_j - g_j) = \mathbf{A + B}$$

$$\text{A} \;=\; V_\xi E_I\left[\sum_t\sum_{i=1}^{m_t} I(y_i=j)+\sum_t\sum_{i=m_t+1}^{N_t} I(\hat{y}_i=j)-\sum_t\sum_{i=1}^{N_t} I(y_i=j)\right]$$

$$= V_\xi\left[\sum_t\sum_{i=1}^{m_t} I(y_i=j)+\sum_t\sum_{i=m_t+1}^{N_t} E_I\left[I(\hat{y}_i=j)\right]-\sum_t\sum_{i=1}^{N_t} I(y_i=j)\right]$$

$$= V_\xi\left[\sum_t\sum_{i=1}^{m_t} I(y_i=j)+\sum_t\sum_{i=m_t+1}^{N_t} \hat{p}_{jt}-\sum_t\sum_{i=1}^{N_t} I(y_i=j)\right]$$

$$= V_\xi\left[\sum_t\sum_{i=1}^{m_t} I(y_i=j)+\sum_t(N_t-m_t)\sum_{i=1}^{m_t}\frac{I(y_i=j)}{m_t}-\sum_t\sum_{i=1}^{N_t} I(y_i=j)\right]$$

$$= V_\xi\left[\sum_t\left[\left(1+\frac{(N_t-m_t)}{m_t}\right)\sum_{i=1}^{m_t} I(y_i=j)-\sum_{i=1}^{N_t} I(y_i=j)\right]\right]$$

$$= V_\xi\left[\sum_t\left[\frac{N_t}{m_t}\sum_{i=1}^{m_t} I(y_i=j)-\sum_{i=1}^{N_t} I(y_i=j)\right]\right]$$

$$= V_\xi\left[\sum_t\left[\frac{N_t}{m_t}\sum_{i=1}^{m_t} I(y_i=j)-\sum_{i=1}^{m_t} I(y_i=j)-\sum_{i=m_t+1}^{N_t} I(y_i=j)\right]\right]$$

$$= V_\xi\left[\sum_t\left[\left(\frac{N_t}{m_t}-1\right)\sum_{i=1}^{m_t} I(y_i=j)-\sum_{i=m_t+1}^{N_t} I(\hat{y}_i=j)\right]\right]$$

$$=\sum_t\left[\left(\frac{N_t}{m_t}-1\right)^2 m_tP_{jt}^{\xi}(1-P_{jt}^{\xi})+\left(N_t-m_t\right)P_{jt}^{\xi}(1-P_{jt}^{\xi})\right]$$ (assuming independence between y_i)

$$= \sum_t\left[\left[\left(\frac{N_t^2-2m_tN_t+m_t^2}{m_t^2}\right)m_t+\left(N_t-m_t\right)\right]P_{jt}^{\xi}(1-P_{jt}^{\xi})\right]$$

$$= \sum_t\left[\left(\frac{N_t^2-N_tm_t}{m_t}\right)P_{jt}^{\xi}(1-P_{jt}^{\xi})\right] \;=\; \sum_t\left[\left(\frac{N_t-m_t}{m_t}\right)N_tP_{jt}^{\xi}(1-P_{jt}^{\xi})\right]$$

$$\text{B} \;=\; E_\xi V_I\left[\sum_t\sum_{i=1}^{m_t} I(y_i=j)+\sum_t\sum_{i=m_t+1}^{N_t} I(\hat{y}_i=j)-\sum_t\sum_{i=1}^{N_t} I(y_i=j)\right]$$

$$= E_\xi V_I\left[\sum_t\sum_{i=m_t+1}^{N_t} I(\hat{y}_i=j)\right] \;=\; E_\xi\left[\sum_t\sum_{i=m_t+1}^{N_t} V_I\left[I(\hat{y}_i=j)\right]\right]$$

$$= E_\xi\left[\sum_t\left(N_t-m_t\right)\hat{p}_{jt}(1-\hat{p}_{jt})\right] \;=\; \sum_t\left(N_t-m_t\right)P_{jt}^{\xi}(1-P_{jt}^{\xi})$$

Thus, $V(\hat{g}_j - g_j)$ = $\sum_t \left[\left(\frac{N_t - m_t}{m_t} \right) N_t P_{jt}^{\xi} (1 - P_{jt}^{\xi}) + (N_t - m_t) P_{jt}^{\xi} (1 - P_{jt}^{\xi}) \right]$

$$= \sum_t \left[(N_t - m_t) \left(\frac{N_t}{m_t} + 1 \right) P_{jt}^{\xi} (1 - P_{jt}^{\xi}) \right]$$

$$= \sum_t \left[\left(\frac{N_t^2 - m_t^2}{m_t} \right) P_{jt}^{\xi} (1 - P_{jt}^{\xi}) \right]$$

Hence, the variance is given by the expression

$$V_{\xi}(\hat{g}_j - g_j) = \sum_t \left[\left(\frac{N_t^2 - m_t^2}{m_t} \right) P_{jt}^{\xi} (1 - P_{jt}^{\xi}) \right] \qquad (2)$$

3.7.2. Some variations in the Probability Distribution Imputation Case

This section introduces a different approach to the model based. In order to compare with the model based approach, the "Finite Population" approach is used in this section. The finite population approach corresponds to the case where no model is used. That is, given a finite population, values are treated as fixed and inferences are based on the distribution of the observed data. This approach is similar to what is known in the literature as "Randomisation Approach" (Little and Rubin, 1987) or "Frequentist Approach" (Rao, 2001) in the sense that they do not require model assistance. However, the finite population approach in this thesis does not make use of sample selection as the other two mentioned.

Additionally to the comparison between the model base and finite population, this approach represents an important aspect in this thesis as simulations in *Chapter 4* and *Chapter 6* are carried out without the use of model, that is, assuming a finite fixed population as in this case.

A difference between this approach and the model based approach presented in previous section is basically that no assumptions are made about the response mechanism. Therefore, bias, variance and variance estimation are assessed with respect to the response mechanism as well as to the imputation methods.

3.7.2.1 Probability Distribution Imputation Case under Finite Population Approach

Let us consider the properties of the estimator proposed for the probability distribution imputation method under the finite population approach. It is important to point out that in this case there are not probabilities involved as the population is fixed. We use the frequency distribution of the observed data to estimate missing quantities.

Bias with respect to the response mechanism

$$E_R E_I\left[\sum_t\sum_{m_t} I(y_i=j)+\sum_t\sum_{m_t+1}^{N_t} I(\hat{y}_i=j)-\sum_t\sum_{N_t} I(y_i=j)\right]$$

$$= E_R\left[\sum_t\sum_{m_t} I(y_i=j)+\sum_t\sum_{m_t+1}^{N_t} E_I\left(I(\hat{y}_i=j)\right)-\sum_t\sum_{N_t} I(y_i=j)\right]$$

$$= E_R\left[\sum_t\sum_{m_t} I(y_i=j)+\sum_t (N_t-m_t)\frac{\sum_{m_t} I(y_i=j)}{m_t}-\sum_t\sum_{N_t} I(y_i=j)\right]$$

$$= E_R\left[\sum_t \frac{N_t}{m_t}\sum_{m_t} I(y_i=j)-\sum_t\sum_{N_t} I(y_i=j)\right]$$

$$= E_R\left[\sum_t \frac{N_t}{m_t}\sum_{N_t} R_i I(y_i=j)-\sum_t\sum_{N_t} I(y_i=j)\right]$$

$$= \sum_t \frac{N_t}{m_t}\sum_{N_t} E_R(R_i) I(y_i=j)-\sum_t\sum_{N_t} I(y_i=j)$$

$$= \sum_t \frac{N_t}{m_t}\frac{m_t}{N_t}\sum_{N_t} I(y_i=j)-\sum_t\sum_{N_t} I(y_i=j) = 0$$

Therefore, we can see that $\hat{g}_j$ is an unbiased estimator of g_j under the finite population approach.

Variance with respect to the response mechanism

$$V_R E_I(\hat{g}_j - g_j) + E_R V_I(\hat{g}_j - g_j) = A + B$$

$$\mathrm{A} = \quad V_R E_I\left[\sum_t\sum_{m_t} I(y_i=j)+\sum_t\sum_{m_t+1}^{N_t} I(\hat{y}_i=j)-\sum_t\sum_{N_t} I(y_i=j)\right]$$

$$= V_R\left[\sum_t\sum_{m_t} I(y_i=j) + \sum_t\sum_{m_t+1}^{N_t} E_I\left(I(\hat{y}_i=j)\right) - \sum_t\sum_{N_t} I(y_i=j)\right]$$

$$= V_R\left[\sum_t\sum_{m_t} I(y_i=j) + \sum_t (N_t-m_t)\frac{\sum_{m_t} I(y_i=j)}{m_t} - \sum_t\sum_{N_t} I(y_i=j)\right]$$

$$= V_R\left[\sum_t \frac{N_t}{m_t}\sum_{m_t} I(y_i=j) - \sum_t\sum_{N_t} I(y_i=j)\right]$$

$$= V_R\left[\sum_t \frac{N_t}{m_t}\sum_{N_t} R_i I(y_i=j) - \sum_t\sum_{N_t} I(y_i=j)\right]$$

$$= V_R\left[\sum_t \frac{N_t}{m_t}\sum_{N_t} R_i I(y_i=j)\right]$$

$$= \sum_t \frac{N_t^2}{m_t^2}\left[\sum_{N_t}\left(I(y_i=j)\right)^2 V_R\left(R_i\right) + \sum_{i\in N_t}\sum_{\substack{i'\in N_t\\ i\neq i'}} I(y_i=j)I(y_{i'}=j)COV_R\left(R_i,R_{i'}\right)\right]$$

$$= \sum_t \frac{N_t^2}{m_t^2}\left[\frac{m_t}{N_t}\left(1-\frac{m_t}{N_t}\right)\sum_{N_t}\left(I(y_i=j)\right)^2\right]$$

$$- \sum_t \frac{N_t^2}{m_t^2}\left[\frac{m_t}{N_t(N_t-1)}\left(1-\frac{m_t}{N_t}\right)\sum_{i\in N_t}\sum_{\substack{i'\in N_t\\ i\neq i'}} I(y_i=j)I(y_{i'}=j)\right]$$

$$= \sum_t\left[\left(\frac{N_t-m_t}{m_t}\right)\sum_{N_t}\left(I(y_i=j)\right)^2 - \left(\frac{N_t-m_t}{m_t(N_t-1)}\right)\sum_{i\in N_t}\sum_{\substack{i'\in N_t\\ i\neq i'}} I(y_i=j)I(y_{i'}=j)\right]$$

$$= \sum_t\left(\frac{N_t-m_t}{m_t}\right)\left[\sum_{N_t}\left(I(y_i=j)\right) - \frac{1}{(N_t-1)}\sum_{i\in N_t}\sum_{\substack{i'\in N_t\\ i\neq i'}} I(y_i=j)I(y_{i'}=j)\right]$$

$$= \sum_t\left(\frac{N_t-m_t}{m_t}\right)\left[N_tP_{jt} - \frac{N_tP_{jt}(N_tP_{jt}-1)}{(N_t-1)}\right]$$

$$= \sum_t\left(\frac{N_t-m_t}{m_t}\right)\left[\frac{N_t(N_t-1)P_{jt} - N_t^2P_{jt}^2 + N_tP_{jt}}{(N_t-1)}\right]$$

$$= \sum_t\left(\frac{N_t-m_t}{m_t}\right)\frac{N_t^2}{(N_t-1)}P_{jt}(1-P_{jt}) \approx \sum_t\left(\frac{N_t-m_t}{m_t}\right)N_tP_{jt}(1-P_{jt})$$

$$\text{B} = \quad E_R V_I \left[\sum_t \sum_{m_t} I(y_i = j) + \sum_t \sum_{m_t+1}^{N_t} I(\hat{y}_i = j) - \sum_t \sum_{N_t} I(y_i = j) \right]$$

$$= E_R V_I \left[\sum_t \sum_{m_t+1}^{N_t} I(\hat{y}_i = j) \right] = E_R \left[\sum_t (N_t - m_t) \hat{p}_{jt} (1 - \hat{p}_{jt}) \right]$$

$$= \sum_t \left(N_t - m_t\right) \frac{1}{m_t^{\,2}} \left[\frac{m_t^{\,2}}{N_t} \sum_{N_t} I(y_i = j) - \frac{m_t}{N_t} \sum_{N_t} I(y_i = j) \right]$$

$$- \sum_t \left(N_t - m_t\right) \frac{1}{m_t^{\,2}} \left[\frac{(m_t - 1)}{N_t (N_t - 1)} \sum_{i \in N_t} \sum_{\substack{i' \in N_t \\ i \neq i'}} I(y_i = j) I(y_{i'} = j) \right]$$

$$= \sum_t \left(N_t - m_t\right) \frac{1}{m_t^{\,2}} \left[m_t (m_t - 1) P_{jt} - \frac{m_t (m_t - 1)}{N_t (N_t - 1)} N_t P_{jt} (N_t P_{jt} - 1) \right]$$

$$= \sum_t \left(\frac{N_t - m_t}{m_t} \right) (m_t - 1) P_{jt} \left[1 - \frac{(N_t P_{jt} - 1)}{(N_t - 1)} \right] = \sum_t \left(\frac{\left(N_t - m_t\right)(m_t - 1) N_t}{m_t \left(N_t - 1\right)} \right) P_{jt} (1 - P_{jt})$$

$$\approx \sum_t \left(N_t - m_t\right) P_{jt} (1 - P_{jt})$$

$$\text{A + B} = \sum_t \left(\frac{N_t - m_t}{m_t} \right) \frac{N_t^2}{N_t - 1} P_{jt} (1 - P_{jt}) + \sum_t \left(\frac{\left(N_t - m_t\right)(m_t - 1) N_t}{m_t \left(N_t - 1\right)} \right) P_{jt} (1 - P_{jt})$$

$$= \sum_t \left(\frac{N_t - m_t}{m_t} \right) \left(\frac{N_t^2}{N_t - 1} + \frac{(m_t - 1) N_t}{N_t - 1} \right) P_{jt} (1 - P_{jt})$$

$$= \sum_t \left(\frac{N_t - m_t}{m_t} \right) \left(\frac{N_t (N_t + m_t - 1)}{N_t - 1} \right) P_{jt} (1 - P_{jt})$$

$$\approx \sum_t \left(\frac{N_t - m_t}{m_t} \right) (N_t + m_t) P_{jt} (1 - P_{jt}) = \sum_t \left(\frac{N_t^{\,2} - m_t^{\,2}}{m_t} \right) P_{jt} (1 - P_{jt})$$

Then, the variance of the difference $(\hat{g}_j - g_j)$ is given by the approximation

$$V_F (\hat{g}_j - g_j) \approx \sum_t \left(\frac{N_t^{\,2} - m_t^{\,2}}{m_t} \right) P_{jt} (1 - P_{jt}) \qquad (3)$$

It can be noticed that the results of the variance obtained in this section (which is an approximation) is the same as the results of the variance obtained in the case of model-based approach (equation 2). Therefore, we can say that the variance of the estimator of the total under the model-based is approximately equal to the variance of the same estimator under the finite population approach.

3.7.2.2 Frequency Distribution Method, another approach to Probability Distribution Imputation Method

Sometimes it is not easy or there is not enough time to implement all the theoretical material explained before in practice. Therefore, an easier way of implementing the probability distribution imputation method in practice is proposed hereafter. This new approach involves certain changes in the way in which values are given to the recipients. We call the new version "Frequency Distribution Method".

In contrast to the probability distribution method, the frequency distribution method does not give probabilities to the recipient to be imputed with a certain category depending on the probability distribution obtained in a specific terminal node. The frequency distribution method imputes all the missing records present in a terminal node by using the frequency distribution of the observed values in that terminal node. That means, the number of records imputed with a specific category j will depend on the percentage of observed records with that category present in that specific terminal node. This new approach makes the application of the procedure easier and faster, facilitating and optimising the use of computational resources.

It can be noticed that the main difference between the Probability Distribution and the Frequency Distribution imputation methods is that, in the Frequency Distribution, the number of records to be imputed in a specific category is fixed as it is based of the frequencies of the response variable in a specific terminal node, while in the Probability Distribution case, the expected value of the number of records to be imputed in a specific category depends on the probability for that category in that specific terminal node, that is, $m_t P^{\xi}_{jt}$.

To illustrate this method, suppose that we have the results, used before in this chapter, of a particular terminal node t of a tree. This terminal node has the following frequency distribution for the variable to be imputed

$$\hat{p}_{jt} = \begin{cases} .108; & j=1 \\ .596; & j=2 \\ .242; & j=3 \\ .054; & j=4 \end{cases}$$

This is, for that specific terminal node, 10.8% of the records of the observed variable have category 1; 59.6% of the records have category 2; 24.2% of the records have category 3; and only 5.4% of the records have category 4.

Then, for the frequency distribution method all the records with missing values for the variable to be imputed that end up in this specific terminal node will be imputed as follows: 10,8% of the records will be imputed as category $j=1$; 59,6% will be imputed as category $j=2$, and so on. The selection of the records for the class assignment is carried out by using a simple random selection procedure without replacement for each category to be imputed.

As this new approach makes the application of the procedure easier and faster, the simulations presented in the next chapter are carried out using this methodology. Therefore, its properties are reviewed hereafter.

Bias with respect to model

Let us calculate the expected value of the difference between the estimator of g_j and the parameter

$$E_{\xi I}\left[\hat{g}_j - g_j\right] = E_{\xi I}\left[\sum_t\sum_{i=1}^{m_t} I(y_i=j) + \sum_t \sum_{i=m_t+1}^{N_t} I(\hat{y}_i = j) - \sum_t\sum_{i=1}^{N_t} I(y_i=j)\right]$$

$$= E_\xi\left[\sum_t\sum_{i=1}^{m_t} I(y_i=j)+\sum_t\sum_{i=m_t+1}^{N_t}\hat{p}_{jt} - \sum_t\sum_{i=1}^{N_t} I(y_i=j)\right]$$

$$= E_\xi\left[\sum_t\sum_{i=1}^{m_t} I(y_i=j)+\sum_t (N_t-m_t)\sum_{i=1}^{m_t}\frac{I(y_i=j)}{m_t} - \sum_t\sum_{i=1}^{N_t} I(y_i=j)\right]$$

$$= E_\xi\left[\sum_t\left[\sum_{i=1}^{m_t} I(y_i=j)+(N_t-m_t)\sum_{i=1}^{m_t}\frac{I(y_i=j)}{m_t}-\sum_{i=1}^{N_t}I(y_i=j)\right]\right]$$

$$=\sum_t\left[E_\xi\left(\sum_{i=1}^{m_t}I(y_i=j)\right)+E_\xi\left((N_t-m_t)\sum_{i=1}^{m_t}\frac{I(y_i=j)}{m_t}\right)-E_\xi\left(\sum_{i=1}^{N_t}I(y_i=j)\right)\right]$$

$$=\sum_t\left[\left(\sum_{i=1}^{m_t}E_\xi\left(I(y_i=j)\right)\right)+\left((N_t-m_t)\left(\sum_{i=1}^{m_t}\frac{E_\xi\left(I(y_i=j)\right)}{m_t}\right)\right)-\left(\sum_{i=1}^{N_t}E_\xi\left(I(y_i=j)\right)\right)\right]$$

$$=\sum_t\left[\left(m_tP_{jt}^{\xi}\right)+\left((N_t-m_t)m_t\frac{P_{jt}^{\xi}}{m_t}\right)-\left(N_tP_{jt}^{\xi}\right)\right] = \sum_t\left[m_tP_{jt}^{\xi}+N_tP_{jt}^{\xi}-m_tP_{jt}^{\xi}-N_tP_{jt}^{\xi}\right] = 0$$

Hence, $\hat{g}_j$ is an unbiased estimator of g_j under the model assumptions.

__Variance with respect to model__

We have $V(\hat{g}_j - g_j) \; = V_\xi E_I(\hat{g}_j - g_j) + E_\xi V_I(\hat{g}_j - g_j) \; = \mathrm{A} + \mathrm{B}$

$$\mathrm{A} \; = \; V_\xi\left[\sum_t \sum_{i=1}^{m_t} I(y_i = j) + \sum_t \sum_{i=m_t+1}^{N_t} I(\hat{y}_i = j) - \sum_t \sum_{i=1}^{N_t} I(y_i = j)\right]$$

$$= V_\xi\left[\sum_t \sum_{i=1}^{m_t} I(y_i = j) + \sum_t (N_t - m_t) \sum_{i=1}^{m_t} \frac{I(y_i = j)}{m_t} - \sum_t \sum_{i=1}^{N_t} I(y_i = j)\right]$$

$$= V_\xi\left[\sum_t \left[\left(1 + \frac{(N_t - m_t)}{m_t}\right) \sum_{i=1}^{m_t} I(y_i = j) - \sum_{i=1}^{N_t} I(y_i = j)\right]\right]$$

$$= V_\xi\left[\sum_t \left[\frac{N_t}{m_t} \sum_{i=1}^{m_t} I(y_i = j) - \sum_{i=1}^{N_t} I(y_i = j)\right]\right]$$

$$= V_\xi\left[\sum_t \left[\frac{N_t}{m_t} \sum_{i=1}^{m_t} I(y_i = j) - \sum_{i=1}^{m_t} I(y_i = j) - \sum_{i=m_t+1}^{N_t} I(y_i = j)\right]\right]$$

$$= V_\xi\left[\sum_t \left[\left(\frac{N_t}{m_t} - 1\right) \sum_{i=1}^{m_t} I(y_i = j) - \sum_{i=m_t+1}^{N_t} I(y_i = j)\right]\right]$$

$$= \sum_t \left[\left(\frac{N_t}{m_t} - 1\right)^2 m_t P_{jt}^{\xi}(1 - P_{jt}^{\xi}) + (N_t - m_t) P_{jt}^{\xi}(1 - P_{jt}^{\xi})\right]$$ (assuming independence between y_i)

$$= \sum_t \left[\left[\left(\frac{N_t^2 - 2m_t N_t + m_t^2}{m_t^2}\right) m_t + (N_t - m_t)\right] P_{jt}^{\xi}(1 - P_{jt}^{\xi})\right]$$

$$= \sum_t \left[\left(\frac{N_t^2 - N_t m_t}{m_t}\right) P_{jt}^{\xi}(1 - P_{jt}^{\xi})\right] \; = \; \sum_t \left[\left(\frac{N_t - m_t}{m_t}\right) N_t P_{jt}^{\xi}(1 - P_{jt}^{\xi})\right]$$

$$\mathrm{B} \; = \; E_\xi V_I\left[\sum_t \sum_{i=1}^{m_t} I(y_i = j) + \sum_t \sum_{i=m_t+1}^{N_t} I(\hat{y}_i = j) - \sum_t \sum_{i=1}^{N_t} I(y_i = j)\right] = 0$$

Then, the variance of the difference between the estimator of g_j and the parameter is given by the expression

$$V_\xi(\hat{g}_j - g_j) = \sum_t \left(\frac{N_t - m_t}{m_t}\right) N_t P_{jt}^{\xi}(1 - P_{jt}^{\xi}) \qquad (4)$$

Bias with respect to the response mechanism

$$E_R\left[\sum_t \sum_{m_t} I(y_i = j) + \sum_t \sum_{m_t+1}^{N_t} I(\hat{y}_i = j) - \sum_t \sum_{N_t} I(y_i = j)\right]$$

$$= E_R\left[\sum_t \sum_{m_t} I(y_i = j) + \sum_t (N_t - m_t) \frac{\sum_{m_t} I(y_i = j)}{m_t} - \sum_t \sum_{N_t} I(y_i = j)\right]$$

$$= E_R\left[\sum_t \frac{N_t}{m_t} \sum_{m_t} I(y_i = j) - \sum_t \sum_{N_t} I(y_i = j)\right]$$

$$= E_R\left[\sum_t \frac{N_t}{m_t} \sum_{N_t} R_i I(y_i = j) - \sum_t \sum_{N_t} I(y_i = j)\right]$$

$$= \sum_t \frac{N_t}{m_t} \sum_{N_t} E_R(R_i) I(y_i = j) - \sum_t \sum_{N_t} I(y_i = j)$$

$$= \sum_t \frac{N_t}{m_t} \frac{m_t}{N_t} \sum_{N_t} I(y_i = j) - \sum_t \sum_{N_t} I(y_i = j) = 0$$

Thus, we can see that $\hat{g}_j$ is an unbiased estimator of g_j under the finite population approach.

Variance with respect to the response mechanism

Let us calculate $V_R E_I(\hat{g}_j - g_j) + E_R V_I(\hat{g}_j - g_j) = A + B$

$$\text{A} = \quad V_R\left[\sum_t \sum_{m_t} I(y_i = j) + \sum_t \sum_{m_t+1}^{N_t} I(\hat{y}_i = j) - \sum_t \sum_{N_t} I(y_i = j)\right]$$

$$= V_R\left[\sum_t\sum_{m_t} I(y_i=j)+\sum_t (N_t-m_t)\frac{\sum_{m_t} I(y_i=j)}{m_t}-\sum_t\sum_{N_t} I(y_i=j)\right]$$

$$= V_R\left[\sum_t \frac{N_t}{m_t}\sum_{m_t} I(y_i=j)-\sum_t\sum_{N_t} I(y_i=j)\right]$$

$$= V_R\left[\sum_t \frac{N_t}{m_t}\sum_{N_t} R_i I(y_i=j)-\sum_t\sum_{N_t} I(y_i=j)\right]$$

$$= V_R\left[\sum_t \frac{N_t}{m_t}\sum_{N_t} R_i I(y_i=j)\right]$$

$$= \sum_t \frac{N_t^2}{m_t^2}\left[\sum_{N_t}\left(I(y_i=j)\right)^2 V_R\left(R_i\right)+\sum_{i\in N_t}\sum_{\substack{i'\in N_t\\ i\neq i'}} I(y_i=j)I(y_{i'}=j)COV_R\left(R_i,R_{i'}\right)\right]$$

$$= \sum_t \frac{N_t^2}{m_t^2}\left[\frac{m_t}{N_t}\left(1-\frac{m_t}{N_t}\right)\sum_{N_t}\left(I(y_i=j)\right)^2\right]$$

$$-\sum_t \frac{N_t^2}{m_t^2}\left[\frac{m_t}{N_t(N_t-1)}\left(1-\frac{m_t}{N_t}\right)\sum_{i\in N_t}\sum_{\substack{i'\in N_t\\ i\neq i'}} I(y_i=j)I(y_{i'}=j)\right]$$

$$= \sum_t\left[\left(\frac{N_t-m_t}{m_t}\right)\sum_{N_t}\left(I(y_i=j)\right)^2-\left(\frac{N_t-m_t}{m_t(N_t-1)}\right)\sum_{i\in N_t}\sum_{\substack{i'\in N_t\\ i\neq i'}} I(y_i=j)I(y_{i'}=j)\right]$$

$$= \sum_t\left(\frac{N_t-m_t}{m_t}\right)\left[\sum_{N_t}\left(I(y_i=j)\right)-\frac{1}{(N_t-1)}\sum_{i\in N_t}\sum_{\substack{i'\in N_t\\ i\neq i'}} I(y_i=j)I(y_{i'}=j)\right]$$

$$= \sum_t\left(\frac{N_t-m_t}{m_t}\right)\left[N_tP_{jt}-\frac{N_tP_{jt}(N_tP_{jt}-1)}{(N_t-1)}\right]$$

$$= \sum_t\left(\frac{N_t-m_t}{m_t}\right)\left[\frac{N_t(N_t-1)P_{jt}-N_t^2P_{jt}^2+N_tP_{jt}}{(N_t-1)}\right]$$

$$= \sum_t\left(\frac{N_t-m_t}{m_t}\right)\frac{N_t^2}{(N_t-1)}P_{jt}(1-P_{jt}) \approx \sum_t\left(\frac{N_t-m_t}{m_t}\right)N_tP_{jt}(1-P_{jt})$$

B = $$E_RV_I\left[\sum_t\sum_{m_t} I(y_i=j)+\sum_t\sum_{m_t+1}^{N_t} I(\hat{y}_i=j)-\sum_t\sum_{N_t} I(y_i=j)\right] = 0$$

A + B $\approx \sum_t \left(\frac{N_t - m_t}{m_t} \right) N_t P_{jt}(1 - P_{jt})$

Then, the variance of $(\hat{g}_j - g_j)$ is approximated by the expression

$$V_F(\hat{g}_j - g_j) = \sum_t \left(\frac{N_t - m_t}{m_t} \right) N_t P_{jt}(1 - P_{jt}) \qquad (5)$$

Again, given the results of the variance obtained for the estimator of the total under model-based in the case of frequency distribution (equation 4), we can see that this is approximately equal to the variance of the same estimator under the finite population approach (equation 5).

3.7.3. *Highest Probability Imputation Case*

As explained before, the second imputation method applies the same value to all the records with $\mathbf{Y}$ missing using the value with highest probability for $\mathbf{Y}$ in the observed data.
Again, the estimator of the total for $\mathbf{Y}$ can be written as follows

$$\hat{g}_j = \sum_r I(y_i = j) + \sum_{nr} I(\hat{y}_i = j) = \sum_t \sum_{r_t} I(y_i = j) + \sum_t \sum_{nr_t} I(j_t^* = j)$$

with j_t^* equal to the modal category in node t.

That is, all the $(N_t - m_t)$ missing values will be replaced by the modal category of that specific terminal node t, j_t^*.

Thus, $\hat{g}_j - g_j = \sum_t \sum_{i=1}^{m_t} I(y_i = j) + \sum_t (N_t - m_t) I(j_t^* = j) - \sum_t \sum_{i=1}^{N_t} I(y_i = j)$

Bias

The expectation of the difference between the estimator of g_j and the parameter is given by

$$E_\xi\left[\hat{g}_j - g_j\right] = E_\xi\left[\sum_t \sum_{i=1}^{m_t} I(y_i = j) + \sum_t (N_t - m_t) I(j_t^* = j) - \sum_t \sum_{i=1}^{N_t} I(y_i = j)\right]$$

$$= E_{\xi}\left[\sum_{t}\sum_{i=1}^{m_t} I(y_i = j)\right] + E_{\xi}\left[\sum_{t}(N_t - m_t) I(j_t^{*} = j)\right] - E_{\xi}\left[\sum_{t}\sum_{i=1}^{N_t} I(y_i = j)\right]$$

For simplicity, we will consider the case when y_i can take only two categories

- The first term of the last equation is

$$E_{\xi}\left[\sum_{t}\sum_{i=1}^{m_t} I(y_i = j)\right] = \sum_{t}\sum_{i=1}^{m_t} E_{\xi}\left[I(y_i = j)\right] = \sum_{t} m_t P_{jt}^{\xi}$$

- The second term can be obtained as follows

$$E_{\xi}\left[\sum_{t}(N_t - m_t) I(j_t^{*} = j)\right] = \sum_{t}(N_t - m_t) E_{\xi}\left[I(j_t^{*} = j)\right]$$

Where

$$\begin{aligned} E_{\xi}\left[I(j_t^{*} = j)\right] &= P_{\xi}(j_t^{*} = j) = P_{\xi}\left(\hat{p}_{jt} \geq \hat{p}_{lt}\right) \text{ for } l \neq j \\ &= P_{\xi}\left(\hat{p}_{jt} \geq 1 - \hat{p}_{jt}\right) = P_{\xi}\left(\hat{p}_{jt} \geq 0.5\right) \\ &= P_{\xi}\left(\frac{\sum_{i=1}^{m_t} I(y_i = j)}{m_t} \geq 0.5\right) = \sum_{a=[m_t/2]}^{m_t} \binom{m_t}{a} \left(P_{jt}^{\xi}\right)^{a} \left(1 - P_{jt}^{\xi}\right)^{m_t - a} \end{aligned}$$

since $\sum_{i=1}^{m_t} I(y_i = j) \sim Bin(m_t, P_{jt}^{\xi})$

- The third term is

$$E_{\xi}\left[\sum_{t}\sum_{i=1}^{N_t} I(y_i = j)\right] = \sum_{t}\sum_{i=1}^{N_t} E_{\xi}\left[I(y_i = j)\right] = \sum_{t} N_t P_{jt}^{\xi}$$

We finally have,

$$\begin{aligned} E_{\xi I}\left[\hat{g}_j - g_j\right] &= \sum_{t} m_t P_{jt}^{\xi} + \sum_{t}(N_t - m_t) \sum_{a=[m_t/2]}^{m_t} \binom{m_t}{a} \left(P_{jt}^{\xi}\right)^{a} \left(1 - P_{jt}^{\xi}\right)^{m_t - a} - \sum_{t} N_t P_{jt}^{\xi} \\ &= \sum_{t}\left[m_t P_{jt}^{\xi} + (N_t - m_t) \sum_{a=[m_t/2]}^{m_t} \left[\binom{m_t}{a} \left(P_{jt}^{\xi}\right)^{a} \left(1 - P_{jt}^{\xi}\right)^{m_t - a}\right] - N_t P_{jt}^{\xi}\right] \\ &= \sum_{t}\left[(N_t - m_t) \sum_{a=[m_t/2]}^{m_t} \left[\binom{m_t}{a} \left(P_{jt}^{\xi}\right)^{a} \left(1 - P_{jt}^{\xi}\right)^{m_t - a}\right] - (N_t - m_t) P_{jt}^{\xi}\right] \end{aligned}$$

$$= \sum_t \left[(N_t - m_t) \left[\sum_{a=[m_t/2]}^{m_t} \left[\binom{m_t}{a} (P_{jt}^{\xi})^a \left(1 - P_{jt}^{\xi}\right)^{m_t - a} \right] - P_{jt}^{\xi} \right] \right]$$

Hence, in general, $\hat{g}_j$ is not an unbiased estimator of g_j under the model assumptions. However, we can see that when m_t is large and P_{jt}^{ξ} is much larger than 0.5 (close to 1), then $P_\xi\left(\hat{p}_{jt} \geq 0.5\right)$ is approximately 1 and the bias become small. That means, the "purer" the terminal nodes are the smaller the bias.
Therefore, the Highest Probability Distribution seems to be a good low biased method when the classification is accurate, otherwise the bias could be very large.

Finally, an expression for estimating the bias in the case in which y_i takes only two different categories can be $\sum_t \left[(N_t - m_t) \left[\sum_{a=[m_t/2]}^{m_t} \left[\binom{m_t}{a} (\hat{p}_{jt})^a \left(1 - \hat{p}_{jt}\right)^{m_t - a} \right] - \hat{p}_{jt} \right] \right]$.

Variance

In order to obtain an expression for the variance, we write

$$V\left[\hat{g}_j - g_j\right] = V_\xi E_I\left[\hat{g}_j - g_j\right] + E_\xi V_I\left[\hat{g}_j - g_j\right] = V_\xi\left[\hat{g}_j - g_j\right]$$

$$V_\xi\left[\hat{g}_j - g_j\right] = V_\xi\left[\sum_t \sum_{i=1}^{m_t} I(y_i = j) + \sum_t (N_t - m_t) I(j_t^* = j) - \sum_t \sum_{i=1}^{N_t} I(y_i = j)\right]$$

$$= V_\xi\left[\sum_t (N_t - m_t) I(j_t^* = j) - \sum_t \sum_{i=m_t+1}^{N_t} I(y_i = j)\right]$$

$$= \sum_t (N_t - m_t)^2 V_\xi\left[I(j_t^* = j)\right] + \sum_t (N_t - m_t) V_\xi\left[I(y_i = j)\right]$$

- The first expression can be obtained as follows

$$V_\xi\left[I(j_t^* = j)\right] = P_\xi(\hat{p}_{jt} \geq \hat{p}_{lt})\left[1 - P_\xi(\hat{p}_{jt} \geq \hat{p}_{lt})\right]$$

- The second expression is

$$V_\xi\left[I(y_i = j)\right] = P_{jt}^{\xi}(1 - P_{jt}^{\xi})$$

Then, we finally have,

$$V\left[\hat{g}_j - g_j\right] = \sum_t (N_t - m_t)^2 P_\xi(\hat{p}_{jt} \geq \hat{p}_{lt})\left[1 - P_\xi(\hat{p}_{jt} \geq \hat{p}_{lt})\right] + \sum_t \left[(N_t - m_t) P_{jt}^\xi (1 - P_{jt}^\xi)\right]$$

$$= \sum_t (N_t - m_t)\left[(N_t - m_t) P_\xi(\hat{p}_{jt} \geq \hat{p}_{lt})\left[1 - P_\xi(\hat{p}_{jt} \geq \hat{p}_{lt})\right] + P_{jt}^\xi (1 - P_{jt}^\xi)\right]$$

where $P_\xi(\hat{p}_{jt} \geq \hat{p}_{lt}) = \sum_{a=[m_t/2]}^{m_t} \binom{m_t}{a} \left(P_{jt}^\xi\right)^a \left(1 - P_{jt}^\xi\right)^{m_t - a}$

3.7.4. Nearest Neighbour Imputation Case

In the case of the nearest neighbour imputation method where the missing value is imputed using the closest donor available, the estimator of the total has the form

$$\hat{g}_j = \sum_t \sum_{r_t} I(y_i = j) + \sum_t \sum_{r_t} A_i I(y_i = j)$$

where A_i is the number of times y_i, $i \in r_t$, is used for imputing a missing record, that is $A_i = \sum_{i \in nr_t} I(d_{i'i} \leq d_{i'l})$ for all $l \in r_t$, with $d_{i'i} = \sum_k I(x_{i'k} \neq x_{ik})$

Thus, $\hat{g}_j - g_j = \sum_t \sum_{i=1}^{m_t} I(y_i = j) + \sum_t \sum_{i=1}^{m_t} A_i I(y_i = j) - \sum_t \sum_{i=1}^{N_t} I(y_i = j)$

Bias

$$E_\xi\left[\hat{g}_j - g_j\right] = E_\xi\left[\sum_t \sum_{i=1}^{m_t} I(y_i = j) + \sum_t \sum_{i=1}^{m_t} A_i I(y_i = j) - \sum_t \sum_{i=1}^{N_t} I(y_i = j)\right]$$

$$= E_\xi\left[\sum_t \sum_{i=1}^{m_t} I(y_i = j)\right] + E_\xi\left[\sum_t \sum_{i=1}^{m_t} A_i I(y_i = j)\right] - E_\xi\left[\sum_t \sum_{i=1}^{N_t} I(y_i = j)\right]$$

- the first term is

$$E_\xi\left[\sum_t \sum_{i=1}^{m_t} I(y_i = j)\right] = \sum_{m_t} m_t P_{jt}^\xi$$

- the second term is

$$E_{\xi}\left[\sum_{t}\sum_{i=1}^{m_t} A_i I(y_i = j)\right] = \sum_{t}\sum_{i=1}^{m_t} A_i E_{\xi}\left[I(y_i = j)\right] = \sum_{t}\sum_{i=1}^{m_t} A_i P_{jt}^{\xi} = \sum_{t}(N_t - m_t)P_{jt}^{\xi}$$

since A_i is a fixed quantity given its definition and the assumption of the model, and each donor is only used for imputing cases within the same terminal node t.

- the third term is

$$E_{\xi}\left[\sum_{t}\sum_{i=1}^{N_t} I(y_i = j)\right] = \sum_{m_t} N_t P_{jt}^{\xi}$$

Then, we finally have

$$\begin{aligned} E_{\xi}\left[\hat{g}_j - g_j\right] &= \sum_{m_t} m_t P_{jt}^{\xi} + \sum_{t}(N_t - m_t)P_{jt}^{\xi} - \sum_{m_t} N_t P_{jt}^{\xi} \\ &= \sum_{t}(N_t - m_t)P_{jt}^{\xi} - (N_t - m_t)P_{jt}^{\xi} = 0 \end{aligned}$$

Therefore, $\hat{g}_j$ is an unbiased estimator of g_j.

Variance

In order to obtain an expression for the variance we have

$$V\left[\hat{g}_j - g_j\right] = V_{\xi}E_I\left[\hat{g}_j - g_j\right] + E_{\xi}V_I\left[\hat{g}_j - g_j\right] = V_{\xi}\left[\hat{g}_j - g_j\right]$$

$$\begin{aligned} V_{\xi}\left[\hat{g}_j - g_j\right] &= V_{\xi}\left[\sum_{t}\sum_{i=1}^{m_t} I(y_i = j) + \sum_{t}\sum_{i=1}^{m_t} A_i I(y_i = j) - \sum_{t}\sum_{i=1}^{N_t} I(y_i = j)\right] \\ &= V_{\xi}\left[\sum_{t}\sum_{i=1}^{m_t} A_i I(y_i = j) - \sum_{t}\sum_{i=m_t+1}^{N_t} I(y_i = j)\right] \\ &= \sum_{t}\sum_{i=1}^{m_t}(A_i)^2 V_{\xi}\left[I(y_i = j)\right] + \sum_{t}(N_t - m_t)V_{\xi}\left[I(y_i = j)\right] \end{aligned}$$

since A_i is a fixed quantity

$$= \sum_{t}\sum_{i=1}^{m_t}(A_i)^2 P_{jt}^{\xi}(1 - P_{jt}^{\xi}) + \sum_{t}(N_t - m_t)P_{jt}^{\xi}(1 - P_{jt}^{\xi})$$

$$= \sum_t \left(N_t - m_t + \sum_{i=1}^{m_t} (A_i)^2 \right) P_{jt}^{\xi} (1 - P_{jt}^{\xi})$$

Let us define $S_{At}^{\,2} = \dfrac{\sum_{i=1}^{m_t} (A_i - \bar{A})^2}{m_t}$ as the variability of A_i, where $\bar{A} = \dfrac{\sum_{i=1}^{m_t} (A_i)}{m_t} = \dfrac{N_t - m_t}{m_t}$ given that $\sum_{i=1}^{m_t} A_i = (N_t - m_t)$, as explained before. Then, $V_\xi \left[\hat{g}_j - g_j \right]$ can be written as

$$\sum_t \left(N_t - m_t + m_t \left(S_{At}^{\,2} + \left(\frac{N_t - m_t}{m_t} \right)^2 \right) \right) P_{jt}^{\xi} (1 - P_{jt}^{\xi}).$$

It can be noticed that the size of the variance of $\hat{g}_j - g_j$ depends on how big the variability of A_i is, that is, on how donors are used. Therefore, $V_\xi \left[\hat{g}_j - g_j \right]$ is smallest when $S_{At}^{\,2}$ is smallest.

3.7.5. *Comparison of the Variance expressions*

Given that we have obtained an expression for the variance in most of the cases (i.e. for the different imputation methods), it could be useful to compare these expressions in order to find out which of the methods produces larger values. In order to do this, a comparison of the expressions for the variances presented in previous sections in made hereafter.

3.7.5.1. Comparison between Probability Distribution and Frequency Distribution imputation methods

In this case, we have that the variance for the Probability Distribution method has the form

$$V_\xi(\hat{g}_j - g_j) = \sum_t \left[\left(\frac{N_t^2 - m_t^2}{m_t} \right) P_{jt}^{\xi} (1 - P_{jt}^{\xi}) \right]$$

while the variance for the Frequency Distribution case has the form

$$V_\xi(\hat{g}_j - g_j) = \sum_t \left(\frac{N_t - m_t}{m_t} \right) N_t P_{jt}^{\xi} (1 - P_{jt}^{\xi})$$

Therefore, we can see that Probability Distribution will always produce larger variances that Frequency Distribution as $\dfrac{N_t^2 - m_t^2}{m_t} > \dfrac{(N_t - m_t) N_t}{m_t}$.

3.7.5.2. Comparison between Probability Distribution and Nearest Neighbour imputation methods

In this case, we have that the expression for the variance for the Probability Distribution method is $V_\xi(\hat{g}_j - g_j) = \sum_t \left[\left(\frac{N_t^2 - m_t^2}{m_t} \right) P_{jt}^\xi (1 - P_{jt}^\xi) \right]$

On the other hand, the variance for the Nearest Neighbour imputation can be expressed in the following way (as in *Section 3.7.4*)

$$V_\xi \left[\hat{g}_j - g_j \right] = \sum_t \left(N_t - m_t + m_t \left(S_{At}^2 + \left(\frac{N_t - m_t}{m_t} \right)^2 \right) \right) P_{jt}^\xi (1 - P_{jt}^\xi)$$

Then, if these two formulas are compared, we can see the Nearest Neighbour variance produces larger values than the Probability Distribution only if $S_{At}^2 > \frac{N_t - m_t}{m_t}$ for every terminal node t. If this condition is presented in just some of the terminal nodes but not all of them, the results do not seem to be that obvious. The same applies to the case in which $S_{At}^2 = \frac{N_t - m_t}{m_t}$ and in the case in which $S_{At}^2 < \frac{N_t - m_t}{m_t}$.

3.7.5.3. Comparison between Frequency Distribution and Nearest Neighbour imputation methods

In this case, we have that the expression for the variance in the case of Frequency Distribution method is $V_\xi(\hat{g}_j - g_j) = \sum_t \left(\frac{N_t - m_t}{m_t} \right) N_t P_{jt}^\xi (1 - P_{jt}^\xi)$.

On the other hand we have the following expression for the Nearest Neighbour imputation

$$V_\xi \left[\hat{g}_j - g_j \right] = \sum_t \left(N_t - m_t + m_t \left(S_{At}^2 + \left(\frac{N_t - m_t}{m_t} \right)^2 \right) \right) P_{jt}^\xi (1 - P_{jt}^\xi)$$

Then, if these two expressions are compared, we can see the Nearest Neighbour variance will always produce larger values than the Frequency Distribution variance, unless the variability of A_i, S_{At}^2, is equal to zero, in which case the two variances are equal.

3.8 VARIANCE ESTIMATION

3.8.1 Probability Distribution Imputation Case

<u>*Variance estimation with respect to the model*</u>

Consider $\hat{V}(\hat{g}_j - g_j)$ = $\sum_t \left[\left(\frac{N_t^2 - m_t^2}{m_t}\right)\hat{p}_{jt}(1-\hat{p}_{jt})\right]$ as an estimator of the variance of $\left(\hat{g}_j - g_j\right)$. In this case $\hat{p}_{jt}$ is obtained using only the observed data.

In order to assess the bias of this variance estimator, we want to examine $E_\xi E_I(\hat{V}(\hat{g}_j - g_j)) = E_\xi E_I\left[\sum_t \left[\left(\frac{N_t^2 - m_t^2}{m_t}\right)\hat{p}_{jt}(1-\hat{p}_{jt})\right]\right]$

Then, $E_\xi E_I\left[\sum_t \left[\left(\frac{N_t^2 - m_t^2}{m_t}\right)\hat{p}_{jt}(1-\hat{p}_{jt})\right]\right]$ = $E_\xi\left[\sum_t \left[\left(\frac{N_t^2 - m_t^2}{m_t}\right)\hat{p}_{jt}(1-\hat{p}_{jt})\right]\right]$

$$= E_\xi\left[\sum_t \left[\left(\frac{N_t^2 - m_t^2}{m_t}\right)\hat{p}_{jt} - \left(\frac{N_t^2 - m_t^2}{m_t}\right)\left(\hat{p}_{jt}\right)^2\right]\right]$$

$$= \sum_t \left[\left(\frac{N_t^2 - m_t^2}{m_t}\right)E_\xi\left(\hat{p}_{jt}\right) - \left(\frac{N_t^2 - m_t^2}{m_t}\right)\left(E_\xi\left(\hat{p}_{jt}\right)^2\right)\right]$$

- The first term of the last equation is

$$E_\xi\left(\hat{p}_{jt}\right) = E_\xi\left(\sum_{i=1}^{m_t}\frac{I(y_i = j)}{m_t}\right) = m_t\frac{P_{jt}^{\xi}}{m_t} = P_{jt}^{\xi}$$

- The second term of the last equation can be solved as follows

$$E_\xi\left[\left(\hat{p}_{jt}\right)^2\right] = E_\xi\left[\left(\sum_{r_t}\frac{I(y_i = j)}{m_t}\right)^2\right] = \frac{1}{m_t^2}\left[E_\xi\left(\sum_{r_t} I(y_i = j)\right)^2\right]$$

$$= \frac{1}{m_t^2}E_\xi\left(\sum_{m_t}\left(I(y_i = j)\right)^2 + \sum_{i \in m_t}\sum_{\substack{i' \in m_t \\ i \neq i'}} I(y_i = j)I(y_{i'} = j)\right)$$

$$= \frac{1}{m_t^2}\left[\sum_{m_t} P_{jt}^{\xi} + \sum_{i\in m_t}\sum_{\substack{i'\in m_t\\ i\neq i'}} \left(P_{jt}^{\xi}\right)^2\right]$$

$$= \frac{1}{m_t^2}\left[m_t P_{jt}^{\xi} + 2\binom{m_t}{2}\left(P_{jt}^{\xi}\right)^2\right] = \frac{1}{m_t^2}\left[m_t P_{jt}^{\xi} + m_t(m_t-1)\left(P_{jt}^{\xi}\right)^2\right]$$

$$= \frac{1}{m_t}\left[P_{jt}^{\xi} + (m_t-1)\left(P_{jt}^{\xi}\right)^2\right]$$

Then, we finally have

$$E_{\xi}E_I(\hat{V}(\hat{g}_j - g_j)) = \sum_t\left[\left(\frac{N_t^2-m_t^2}{m_t}\right)E_{\xi}\left(\hat{p}_{jt}\right) - \left(\frac{N_t^2-m_t^2}{m_t}\right)E_{\xi}\left[\left(\hat{p}_{jt}\right)^2\right]\right]$$

$$= \sum_t\left[\left(\frac{N_t^2-m_t^2}{m_t}\right)P_{jt}^{\xi} - \left(\frac{N_t^2-m_t^2}{m_t}\right)\frac{1}{m_t}\left[P_{jt}^{\xi} + (m_t-1)\left(P_{jt}^{\xi}\right)^2\right]\right]$$

$$= \sum_t\left[\left(\frac{N_t^2-m_t^2}{m_t}\right)\left[P_{jt}^{\xi} - \frac{1}{m_t}\left[P_{jt}^{\xi} - (m_t-1)\left(P_{jt}^{\xi}\right)^2\right]\right]\right]$$

$$= \sum_t\left[\left(\frac{N_t^2-m_t^2}{m_t}\right)P_{jt}^{\xi}\left[1 - \frac{1}{m_t} - \frac{(m_t-1)}{m_t}P_{jt}^{\xi}\right]\right]$$

$$= \sum_t\left[\left(\frac{N_t^2-m_t^2}{m_t}\right)\frac{(m_t-1)}{m_t}P_{jt}^{\xi}(1-P_{jt}^{\xi})\right]$$

Therefore, an unbiased estimator of $V_{\xi}(\hat{g}_j - g_j)$ is $\sum_t\left[\left(\frac{N_t^2-m_t^2}{m_t-1}\right)\hat{p}_{jt}(1-\hat{p}_{jt})\right]$

Variance estimation with respect to the response mechanism

Let us consider $\hat{V}_F \approx \sum_t\left(\frac{N_t-m_t}{m_t}\right)\left(N_t+m_t\right)\hat{p}_{jt}(1-\hat{p}_{jt})$ as an approximate estimator of the V_F. Then, in order to assess the bias of this variance estimator we want to examine the expected value given the response mechanism and the imputation process of $E_R E_I(\hat{V}_F - V_F)$.

Using previous results, we have that $E_R\left[\hat{p}_{jt}(1-\hat{p}_{jt})\right] \approx P_{jt}(1-P_{jt})$. That makes

$$E_R\left[\sum_t\left(\frac{N_t-m_t}{m_t}\right)(N_t+m_t)\hat{p}_{jt}(1-\hat{p}_{jt})\right] \approx \sum_t\left(\frac{N_t-m_t}{m_t}\right)(N_t+m_t)P_{jt}(1-P_{jt})$$

Then, we have proof that $E_R\left[\hat{V}_F\right]-V_F \approx 0$, that is, $\hat{V}_F$ is an approximately unbiased estimator of V_F.

It can be noticed that the variances, and therefore the estimator for the variances, are approximately equal in both approaches, model-based and finite population approach. Therefore, we can say that the estimator obtained under the model-based approach is unbiased under the finite population approach and vice-versa.

3.8.2 Frequency Distribution Imputation Case

Estimation of the variance with respect to model

Consider $\hat{V}(\hat{g}_j-g_j)$ = $\sum_t\left[\left(\frac{N_t-m_t}{m_t}\right)N_t\hat{p}_{jt}(1-\hat{p}_{jt})\right]$ as an estimator of the variance of $\left(\hat{g}_j-g_j\right)$. In this case $\hat{p}_{jt}$ is obtained using only the observed data.

In order to assess the bias of this variance estimator, we want to examine

$$E_\xi E_I(\hat{V}(\hat{g}_j-g_j)) = E_\xi\left[\sum_t\left[\left(\frac{N_t-m_t}{m_t}\right)N_t\hat{p}_{jt}(1-\hat{p}_{jt})\right]\right]$$

$$= E_\xi\left[\sum_t\left[\left(\frac{N_t-m_t}{m_t}\right)N_t\hat{p}_{jt} - \left(\frac{N_t-m_t}{m_t}\right)N_t\left(\hat{p}_{jt}\right)^2\right]\right]$$

$$= \sum_t\left[\left(\frac{N_t-m_t}{m_t}\right)N_tE_\xi\left(\hat{p}_{jt}\right) - \left(\frac{N_t-m_t}{m_t}\right)N_tE_\xi\left(\hat{p}_{jt}\right)^2\right]$$

- The first term of the last equation is

$$E_\xi\left(\hat{p}_{jt}\right) = E_\xi\left(\sum_{i=1}^{m_t}\frac{I(y_i=j)}{m_t}\right) = m_t\frac{P_{jt}^\xi}{m_t} = P_{jt}^\xi$$

- The second term of the last equation can be solved as follows

$$E_\xi\left(\hat{p}_{jt}\right)^2 = E_\xi\left[\left(\sum_{r_i}\frac{I(y_i=j)}{m_t}\right)^2\right] = \frac{1}{m_t^2}E_\xi\left(\sum_{r_i}I(y_i=j)\right)^2$$

$$= \frac{1}{m_t^2}E_\xi\left(\sum_{m_t}\left(I(y_i=j)\right)^2 + \sum_{i\in m_t}\sum_{\substack{i'\in m_t\\ i\neq i'}} I(y_i=j)I(y_{i'}=j)\right)$$

$$= \frac{1}{m_t^2}\left[\sum_{m_t}P_{jt}^{\xi} + \sum_{i\in m_t}\sum_{\substack{i'\in m_t\\ i\neq i'}}\left(P_{jt}^{\xi}\right)^2\right]$$

$$= \frac{1}{m_t^2}\left[m_tP_{jt}^{\xi} + 2\binom{m_t}{2}\left(P_{jt}^{\xi}\right)^2\right] = \frac{1}{m_t^2}\left[m_tP_{jt}^{\xi} + m_t(m_t-1)\left(P_{jt}^{\xi}\right)^2\right]$$

$$= \frac{1}{m_t}\left[P_{jt}^{\xi} + (m_t-1)\left(P_{jt}^{\xi}\right)^2\right]$$

Then, we finally have

$$E_\xi(\hat{V}(\hat{g}_j-g_j)) = \sum_t\left[\left(\frac{N_t-m_t}{m_t}\right)N_tE_\xi\left(\hat{p}_{jt}\right) - \left(\frac{N_t-m_t}{m_t}\right)N_tE_\xi\left(\hat{p}_{jt}\right)^2\right]$$

$$= \sum_t\left[\left(\frac{N_t-m_t}{m_t}\right)N_tP_{jt}^{\xi} - \left(\frac{N_t-m_t}{m_t}\right)\frac{N_t}{m_t}\left[P_{jt}^{\xi} + (m_t-1)\left(P_{jt}^{\xi}\right)^2\right]\right]$$

$$= \sum_t\left[\left(\frac{N_t-m_t}{m_t}\right)N_t\left[P_{jt}^{\xi} - \frac{1}{m_t}\left[P_{jt}^{\xi} - (m_t-1)\left(P_{jt}^{\xi}\right)^2\right]\right]\right]$$

$$= \sum_t\left[\left(\frac{N_t-m_t}{m_t}\right)N_tP_{jt}^{\xi}\left[1 - \frac{1}{m_t} - \frac{(m_t-1)}{m_t}P_{jt}^{\xi}\right]\right]$$

$$= \sum_t\left[\left(\frac{N_t-m_t}{m_t}\right)\frac{N_t(m_t-1)}{m_t}P_{jt}^{\xi}(1-P_{jt}^{\xi})\right]$$

Therefore, an unbiased estimator of $V_\xi(\hat{g}_j - g_j)$ is given by $\sum_t\left[\left(\frac{N_t-m_t}{m_t-1}\right)N_t\hat{p}_{jt}(1-\hat{p}_{jt})\right]$

Variance estimation with respect to the response mechanism

Given that $V_F \approx \sum_t \left(\frac{N_t - m_t}{m_t} \right) N_t P_{jt} (1 - P_{jt})$, let us consider

$\hat{V}_F \approx \sum_t \left(\frac{N_t - m_t}{m_t} \right) N_t \hat{p}_{jt} (1 - \hat{p}_{jt})$ as an estimator of the variance of $(\hat{g}_j - g_j)$.

Using previous results we have $E_R \left[\hat{p}_{jt} (1 - \hat{p}_{jt}) \right] \approx P_{jt} (1 - P_{jt})$, therefore,

$$E_R \left[\hat{V}_F \right] = E_R \left[\sum_t \left(\frac{N_t - m_t}{m_t} \right) N_t \hat{p}_{jt} (1 - \hat{p}_{jt}) \right] \approx \sum_t \left(\frac{N_t - m_t}{m_t} \right) N_t \hat{p}_{jt} (1 - \hat{p}_{jt}) = V_F$$

Then, we have proof that $\hat{V}_F$ is an approximate unbiased estimator of V_F

As in the case of probability distribution method, it can be noticed that the variances, and therefore the estimator for the variances, are approximately equal in both approaches, model-based and finite population. Therefore, we can say that the estimator obtained under the model-based approach is unbiased under the finite population approach and vice-versa.

3.8.3 Highest Probability Imputation Case

In the estimation of the variance of the difference between the estimator and the parameter we could substitute P_{jt}^{ξ} by the observed proportion of cases with category j in the data, $\hat{p}_{jt}$, in order to obtain a value for that estimation, however, since this method is not an unbiased procedure and the bias appear to be not ignorable, an estimation of the variance seems not to be of major interest.

3.8.4 Nearest Neighbour Imputation Case

Consider $\hat{V}(\hat{g}_j - g_j) = \sum_t \left(N_t - m_t + \sum_{i=1}^{m_t} (A_i)^2 \right) \hat{p}_{jt} (1 - \hat{p}_{jt})$ as an estimator of the variance of $\left(\hat{g}_j - g_j \right)$.

In order to assess the bias of this variance estimator, we want to examine

$$E_\xi(\hat{V}(\hat{g}_j - g_j)) = E_\xi\left[\sum_t \left(N_t - m_t + \sum_{i=1}^{m_t}(A_i)^2\right)\hat{p}_{jt}(1-\hat{p}_{jt})\right]$$

Then, $E_\xi\left[\sum_t \left(N_t - m_t + \sum_{i=1}^{m_t}(A_i)^2\right)\hat{p}_{jt}(1-\hat{p}_{jt})\right]$ = $E_\xi\left[\sum_t \left(N_t - m_t + \sum_{i=1}^{m_t}(A_i)^2\right)\hat{p}_{jt}(1-\hat{p}_{jt})\right]$

= $\sum_t \left(N_t - m_t + \sum_{i=1}^{m_t}(A_i)^2\right)\left[E_\xi\left[\hat{p}_{jt}\right] - E_\xi\left[\hat{p}_{jt}\right]^2\right]$

= $\sum_t \left(N_t - m_t + \sum_{i=1}^{m_t}(A_i)^2\right)\left[P_{jt}^{\xi} - \frac{1}{m_t}\left[P_{jt}^{\xi} - (m_t - 1)\left(P_{jt}^{\xi}\right)^2\right]\right]$ using previous results

= $\sum_t \left(N_t - m_t + \sum_{i=1}^{m_t}(A_i)^2\right)\frac{(m_t - 1)}{m_t} P_{jt}^{\xi}\left(1 - P_{jt}^{\xi}\right)$

Then, an unbiased estimator of $V(\hat{g}_j - g_j)$ is $\sum_t \left(N_t - m_t + \sum_{i=1}^{m_t}(A_i)^2\right)\frac{m_t}{(m_t - 1)}\hat{p}_{jt}\left(1 - \hat{p}_{jt}\right)$

Several papers have been published in the last years about Nearest Neighbour imputation. Many of these publications include the assessment of bias, variance and variance estimation for this case. Rao 2001; Steel and Fay 1995; Fay 1999; Chen and Shao 1999, 2000 and 2001 and Rancourt 1999 are some of the most recent works published in the area of variance estimation.

Rancourt 1999, for example, presents in his paper an estimator for the variance using model based approach. He starts by defining GEIS (Generalised Edit and Imputation System), which is an imputation system for economics surveys developed by Statistics Canada. This system includes the use of nearest neighbour as one of the imputation procedures.
After explaining different surveys in which this system is used, he presents some properties for the nearest neighbour imputation method. First, the imputation procedure is represented by a model. Then bias and variance are estimated.

In this paper, Rancourt denotes the imputation value as follows

$$\hat{y}_k = y_{l(k)} = Bz_{l(k)} + E_{l(k)}$$

where $y_{l(k)}$ is the donor l for unit k and $z_{l(k)}$ is the auxiliary information defining the nearest donor, B is the parameter of the model and $E_{l(k)}$ the error.

The paper follows the decomposition in Sarndal (1992) where $(\hat{Y}_{\bullet s} - Y_U) = (\hat{Y}_s - Y_U) + (\hat{Y}_{\bullet s} - \hat{Y}_s)$ with Y_U as the population total, $\hat{Y}_s$ the estimation of the total given the sample and $\hat{Y}_{\bullet s}$ the estimator of the total given the sample in presence of nonresponse. Therefore, he represents the total variance as $V_{TOL} = V_{SAM} + V_{IMP} + V_{MIX}$.

In this paper, Rancourt presents the imputation variance component obtained in Forget (1999) using the same model presented before. This component has the form

$$V_{IMP} = \frac{N^2}{n^2}\left[\left(\sum_r t_l^2 z_l + \sum_o z_k\right)\sigma^2 + B^2 \sum_o \Delta_k\right]$$

where $\frac{N^2}{n^2}$ is the sampling weight, t_l represents the number of times each donor is used, z_l represents auxiliary information, $B^2 \sum_o \Delta_k$ is the conditional bias of the estimator of the total.

After exposing that the second component of the variance is small and of lower magnitude than the left component, he ends up with an estimator for the variance for the imputation part as follows

$$\hat{V}_{IMP} = \frac{N^2}{n^2}\left(\sum_r t_l^2 z_l + \sum_o z_k\right)\sigma^2$$

We can see that the estimator of the variance obtained in this thesis in the case of nearest neighbour imputation $\sum_t \left(N_t - m_t + \sum_{i=1}^{m_t} (A_i)^2\right) \frac{m_t}{(m_t - 1)} \hat{p}_{jt}\left(1 - \hat{p}_{jt}\right)$ corresponds to the estimator of the variance presented in this paper by Rancourt.

The term $\frac{N^2}{n^2}$ does not apply to our formula since we do not use sampling, $\sum_r t_l^2 z_l$ is the number of times the same donors are used multiply by a variable which in our case would be variable that indicates if the records belongs to a specific terminal node ($\sum_t \sum_{i=1}^{m_t} (A_i)^2$ in our case), $\sum_o z_k$ would be the number of missing cases ($\sum_t (N_t - m_t)$in our case) and σ^2, in our case, is the variance of a binomial given that we are estimating the total of cases which belong to a specific category.

CHAPTER 4

UNIVARIATE CASE SIMULATION

4.1. INTRODUCTION

The aim of this chapter is to assess the performance of using classification trees to form imputation classes, using a specific methodology called CART, which includes the use of a software package especially created for this technique. In order to assess this performance, several simulations were carried out using a database that contains synthetic missing values. In these simulations different classification tree sizes as well as different imputation methods were used in order to compare their effect on results.

Moreover, biases and variances (and expected variance estimators for some cases) were calculated in order to evaluate the properties of the estimators used as explained in *Section 3.7* in *Chapter 3.*

4.2. SIMULATION PROCEDURE

A brief description of all the steps taken in these simulations will be given in this section. This includes the use of the trees, imputation and more general material as, for example, the generation of the synthetic database. A more extended description of each step will be given in the following sections.

4.2.1. Generation of the synthetic database. A synthetic database was created as shown in *Figure 4.3.2.1.* This contains artificial holes for which the real values are known. The holes were created by an ignorable missing mechanism. In this case, the data is considered at least as missing at random.

The generation of the database includes several steps:

a) The first step was to get the database ready for the simulation. As will be explained in the data description, the information used corresponded to the UK 1991 Census. This database was given in an ASCII format and it had to be converted into a readable format such as FOXPRO or SAS for carrying out the simulation. A process in SAS was followed to match the structure of the database (dictionary) with the data itself. The resulting database is called an "original database".

b) Second, once the original database was in a readable format, personal information was separated from household information in order to treat the two sets of variables at different times.

c) The third stage was to find the pattern of missing information in order to create the artificial holes for the comparison. This stage included the elaboration of a SAS routine for finding all the possible combinations of missing information in the database and how much of the total they represented. The output of the routine was a complete list of all possible combinations of missing variables with their corresponding percentage with respect to the total, as shown in *Appendix 1*.

d) Fourth, all the records with missing information were deleted from the database in order to create a "complete database", which is a database with only fully observed records. Only 10.82% of the records were deleted due to the percentage of missing information present in the original database as explained in *Section 4.3.2*.

e) Fifth, after the complete database had been constructed, the pattern of missing information found in c) was used for creating artificial holes. Then, a "synthetic database" was generated by replicating the pattern of missing information randomly on the complete database. This synthetic database therefore contains holes for which the real values are known in order to measure the accuracy of the imputation results.

4.2.2. Growing trees. Different trees were grown for each target variable using the complete database.

a) For each target variable, three different tree-sizes were used in the analysis in order to compare the effect of the size on the imputation results. The selection of the sizes is explained in *Section 4.5.2*.

b) After all the trees had been grown, the records with missing values in the target variables were dropped into each tree to find out which terminal node they will end up in for the imputation. This was made for each different tree-sizes. The complete process is explained in *Section 4.5.3*.

4.2.3. ***Imputing.*** After the different trees were grown, imputation was carried out independently for each of the trees.

a) The three different imputation methods described in *Section 3.4* were combined with the three different tree sizes to obtain 9 different imputation results for each target variable. This was made using trees grown with the complete database.

b) For each of the trees, the imputations were produced independently into each terminal node. Then, the results were summarised in order to be compared with the results from other trees.

4.2.4. Evaluation. Different graphs, tests, biases and variances were used for the evaluation of the imputation.

a) Cross-tabulations between the imputed values and the real values were made for all of the possible combinations of tree sizes and imputation methods.

b) Different graphs were made for all of the above tables in order to compare preservation of joint and marginal distributions and preservation of individual values.

c) Tests were also run for each of the cross-tabulations in order to confirm the preservation of joint and marginal distributions and of individual values.

d) Biases and variances were estimated for imputed variables in order to assess the properties of the estimators used. Additionally, estimation for the variances in the case of Frequency Distribution and Nearest Neighbour were also obtained.

4. 3. DATA

4.3.1. Data description

The database used for the analysis consists of a group of variables measured for one single County of England in the 1991 UK Census. The variables used refer to persons in households. Neither the household variables nor identification variables were included in the analysis.

Because the database was not edited completely for all of the variables and all of the persons, the information used in the analysis is only composed of the variables for which all the records were 100% edited in the database. That is, for all of the persons in the database, all the variables were 100% corrected by the editing process. This is important at the imputation stage as it implies that one can be practically sure that the data do not contain inconsistencies.

Different stages were followed in order to get the database ready for the analysis.
As mentioned before, the first stage was to transform the ASCII database into a readable format as DBASE or SAS file. This included matching the structure of the database (the dictionary) with the database itself, identifying all the variables for all of the records.
The size of the database used (original database) is 222872 records with 23 variables. Because not all of the variables were useful for the analysis, all the identification variables were dropped as well as the variables for which the information was not relevant.

Table 4.3.1.1 shows the final list of variables used for the analysis and their descriptions.

Table 4.3.1.1
List of variables included in the analysis

Variable	Definition
AGE	Age of the person, calculated from date of birth
ALWPRIM	Primary activity last week
COB	Country of birth
ETHNIC	Ethnic origin
LTILL	Long term illness
MARCON	Marital status
SEX	Sex
WELSH	Welsh language abilities

As can be seen, all of the variables are categorical, except for the variable AGE, which is numerical. This variable was converted to a categorical one by grouping it for the analysis.

Because some of the variables originally had too many categories for growing trees, they were collapsed. Then, the criterion was to collapse all the variables with more than ten categories. A complete list of variables with their original categories as well as new ones is Tables 4.3.1.2.

Tables 4.3.1.2
Single Variable Definitions

Table A

AGE

Group 1

Group	New Code
0-4	1
5-9	2
10-15	3
16-18	4
19-21	5
22-24	6
25-29	7
30-34	8
35-39	9
40-44	10
45-49	11
50-54	12
55-59	13
60-64	14
65-69	15
70-74	16
75-79	17
80-84	18
85 +	19

AGE

Group 2

Group	New Code
0-4	7
5-15	6
16-24	5
25-34	4
35-54	3
55-64	2
65 +	1

Table B
ETHNIC

Group	Codes	New Code
White	00	1
Any black including mixed	01 / 02 / 70-80	2
Asian	03-05	3
China / Other including other mixed	06 / 81-97	4

Table C
COUNTRY OF BIRTH

Countries	Codes	New Code
UK	601-609	1
Europe / USA	610-612 / 639-641 / 645-671 / 679	2
Indian Sub-continent	632-635	3
Africa / Caribbean	613-631 / 642-644 / 672-678 / 680	4
Asia / Central and South America / Other	636-638 / 681-702	5

Table D
PRIMARY ACTIVITY LAST WEEK

Primary Activity	Codes	New Codes
Employee working full time /	01 /	1
Employee working part time /	02 /	
Self employed, employing others /	03 /	
Self employed, not employing others /	04 /	
Government employment or training scheme	05	
Waiting to take/start a job /	06 /	2
Unemployed / looking for a work /	07	
At school or in full time education /	08 /	3
Unable to work because of long term disability /	09 /	
Retired from paid work /	10 /	
Looking after home/family /	11 /	
Other economically inactive	12	
No code required	$	4

Table E
SEX

Male	1
Female	2

Table F
WELSH

Does not know Welsh	0
Can speak Welsh	1
Can read Welsh	2

Table G
LONG TERM ILLNESS

Has a health problem	1
Does not have a health problem	2

Table H
MARITAL STATUS

Single	1
Married (first marriage)	2
Remarried	3
Divorced	4
Widowed	5

It is important to point out that new versions of CART are available, handling much more categories for the variables used. However, using too many variables with too many categories can make the process of growing a tree very slow and make the analysis more difficult.

4.3.2. Pattern of missing information

The second stage in the process involved looking at the pattern of missing information present in the data. This stage included the elaboration of a SAS routine for finding all the possible combinations of missing information in the database and how much of the total they represented. The output of this routine was a complete list of all possible combinations of missing variables with their correspondent percentages with respect to the total, as shown in *Appendix 1.*

As can be seen in *Appendix 1* the pattern of missing information was not a straightforward one. This included a large number of combinations (168 combinations in total), with up to 6 different variables missing at the same time. This fact made the possibility of creating a tree for every single combination of missing variables very difficult.

Table 4.3.2.1 shows an example of the missing combinations used in the simulations carried out in this thesis

Table 4.3.2.1

Missing combinations used for the simulations (original database)

COB	ETHNIC	LTILL	Total	Percentage.
	■		3916	16.24
		■	3224	13.37
■			1751	7.26
	■	■	521	0.74
■		■	252	1.04
■	■		178	2.16
■	■	■	520	2.16

It can be seen from this table that with only three variables involved we have seven different combinations of missing information.

The total number of records with missing information is 24116, which represents 10.82 % of the original database (222872 records).

To generate the synthetic database, firstly, all records with at least one missing value were deleted from the original database, obtaining a new "complete database" which contains 198756 records (222872 - 24116). Secondly, the pattern of missing information found at the beginning was randomly reproduced on the complete database. This procedure was carried out using a SAS routine for generating the artificial holes using a simple random sample without replacement. That is, for each combination of variable with missing information, a simple random sample without replacement was selected from the complete database in order to delete their values. The size of each random sample depended on the size of the combination missing, as presented in the last table. In this way, the synthetic database was created, containing 198756 records of which 21520 have missing values (10.827% out of 198756).

Table 4.3.2.2 shows the combinations and their totals used in this analysis after creating the synthetic database

Table 4.3.2.2

Missing combinations used for the simulations (synthetic database)

COB	ETHNIC	LTILL	Total	Percentage.
	■		3492	16.24
		■	1561	13.37
■			2875	7.26
	■	■	159	0.74
■		■	465	1.04
■	■		225	2.16
■	■	■	464	2.16

The sizes of all these databases are shown in *Table 4.3.2.3*

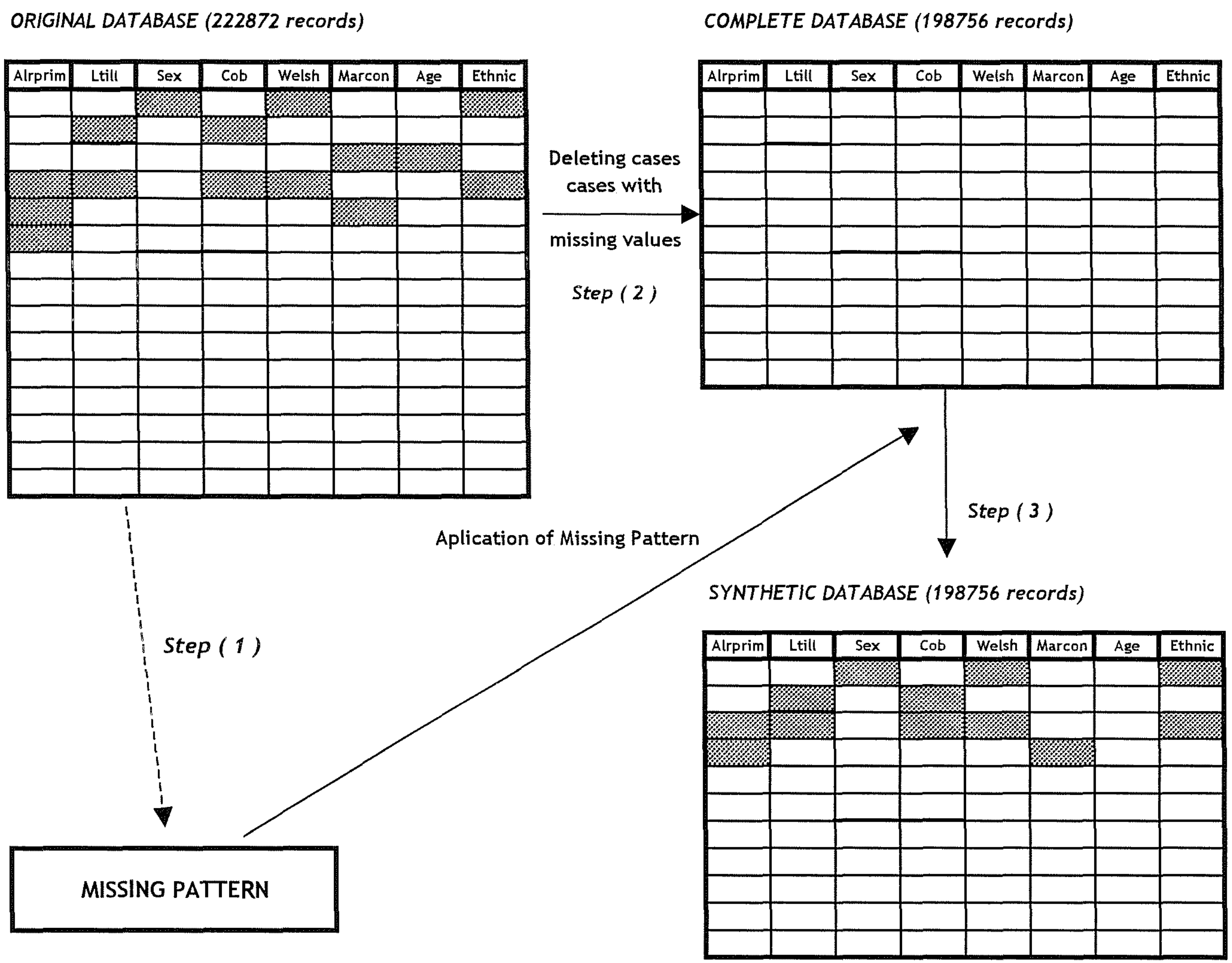
Table 4.3.2.3

Databases sizes and Percentages of missing information

Database	Size	Complete Information	Missing Information
Original Database	222872	198756	10.820%
Complete Database	198756	198756	None
Synthetic Database	198756	177236	10.827%

The whole procedure of the generation of the database used for the analysis is shown in *Figure 4.3.2.1.*

Figure 4.3.2.1: SYNTHETIC DATABASE GENERATION



In order to compare the imputation results, a synthetic database was created (database with artificial holes for which the real values are known)
The steps for the generation of this database is as follows:

Steps

1. Finding the pattern of missingness
Find the pattern of missing information from the original database

2. Creating a complete data
Delete all the records with missing information from the original database

3. Creating synthetic data
Replay the pattern of missing information found in the original database on the complete database

4.3.3. Databases used in the analysis

Every time a combination of variables with missing information is chosen to define a target variable, the remaining information changes as well since different variables are left as covariates (auxiliary variables). Therefore, depending on the combination used as a target, the databases used for the analysis (growing trees, etc.) are different.

The sizes of the three databases depending on the target combination studied for the univariate case are shown in the next table

Table 4.3.3.1

Databases sizes for the Univariate Case

Variable	Database Size	Missing Information
Any	198756 (observed records)	None
Ethnic	198756 - 3916 (Records with Ethnic missing)	Independent variables
Cob	198756 - 1751 (Records with Cob missing)	Independent variables
Ltill	198756 - 3224 (Records with Ltill missing)	Independent variables

It is important to point out that, for simplicity, only one database is used for growing the tree independently of the missing combination study. The database used is the one containing only fully observed records for all the variables and it is the "complete database" shown in Table 4.3.2.3.

It is also important to point out that the analysis could be done including missing information for the covariates, however, for simplicity, only complete information is included in the generation of the tree. A previous study by Mesa, Tsai and Chambers (2000) shows that the inclusion of missing information for growing the tree seems to have no impact on the results when using the same imputation procedures used in this thesis. This study presents the case in which missing information for the auxiliary variables is used in the growing-tree process. In this case, trees were created using and not using missing information for the auxiliary variables in order to compare the impact of the use of missing covariates on the imputation results. The results showed that the use of variables with missing information when creating the classification tree by using CART does not have a major impact on the imputation results. That is, the results obtained for all of the imputation methods are very similar in both cases.

4.4. CLASSIFICATION

The first step of the process is the classification of the units (persons) into the terminal nodes of the tree. This classification was made using the CART methodology described in *Chapter 2.* The specific features of this methodology used for this simulation are now described.

4.4.1. Splitting criterion

Since the variables used are categorical, the criterion based on the Gini index used to classify categorical variables was selected to split the elements in this study.

The impurity function defined by Gini has the form $\delta\left(t\right)=1-SQ$ in which SQ is the sum of squared probabilities $p(j\,|\,t)$. That is, $\delta\left(t\right)=1-\sum_{j} p^{2}\left(j\,|\,t\right)$, where $\delta\left(t\right)$ is a node impurity function defined as $\Phi(p(1\,|\,t),.......,p(J\,|\,t))$ (relative proportion of class j cases in node t), and $p(j\,|\,t)$ is defined by $p(j\,|\,t)=p(j,t)/\,p(t)$, with $p(t)$ is the probability that any case falls into node t where $p(t)=\sum_{j} p(j,t)$

In this analysis the misclassification cost remained constant. In this way, it was assumed that all the costs for misclassifying class j_1 as a class j_2 are equal to 1 for all $j_1 \neq j_2$.

4.4.2. Class assignment rule

Given that one of the imputation methods used involves the use of the class assignment, it is important to define how this assignation was made.

Each terminal node $t \in \widetilde{T}$ has an assigned class $j \in \{1,.....,J\}$, denoted by $j(t)$. This depends on how the prior probabilities are set. In this analysis, the prior probabilities were assumed to be equal, then the class assignment rule $j(t)$ is defined by the plurality rule. That is, node t is classified as the class for which $M_j(t)$ (number of elements with category j in node t in the learning sample) is largest. (See *Section 2.3.4.* for details).

4.5. TREE PROCESS

4.5.1. Growing the tree

Once the variable to be imputed is selected, a tree for that variable is grown. The process of growing a tree is very straightforward. The only necessary requirement for growing a tree using the CART software is the specification of the set of explanatory variables and identification of the response variable. Then, all the software instructions were followed in order to obtain the tree.

When missing values are present in the target variable (variable for which the tree is grown), all the cases are automatically deleted by the software. A case cannot be classified if it does not have its respective class. However, when missing values are present in the auxiliary variables, the cases are still usable for the process of growing the tree. This is possible by the use of surrogates defined by the explanatory variables with non-missing values as explained in *Section 2.3.5* in *Chapter 2*. However, due time constraints, and since including missing information in the auxiliary variable does not seem to have any major impact on the results, databases using missing information in the auxiliary variables are not used in this simulation. An example of using surrogates for classifying missing information in the auxiliary variables is presented by Mesa, Tsai and Chambers (2000).

4.5.2. Selection of the tree size

On occasion, trees can have a very large number of terminal nodes. When the size of the tree is very large, the imputation process becomes a very long and time consuming one. Hence, a decision about the size of the tree used for the analysis is an important aspect to consider in this chapter in order to perform all the simulations required for the analysis. However, the use of different sizes of trees is also desirable in order to compare the effectiveness of the imputation procedures when different numbers of terminal nodes (imputation classes) are used. The process followed for making those decisions is given hereafter.

When a tree is very large, it is usually necessary to find ways to "prune" it without compromising its effectiveness. One of the most common ways to prune a classification tree is by the use of its misclassification rate. As explained in *Section 2.3.6* in *Chapter 2*, this rate is a measure of the percentage of cases misclassified by the *class assignment rule* used in any terminal node. As noted before, each terminal node is given a class, in this case, depending on the modal category for that node.

Figure 4.5.2.1 plots the change in the misclassification rate calculated by the cross-validation method of a CART tree mentioned in *Section 2.3.6* for the variable Primary Activity Last Week by the number of terminal nodes.

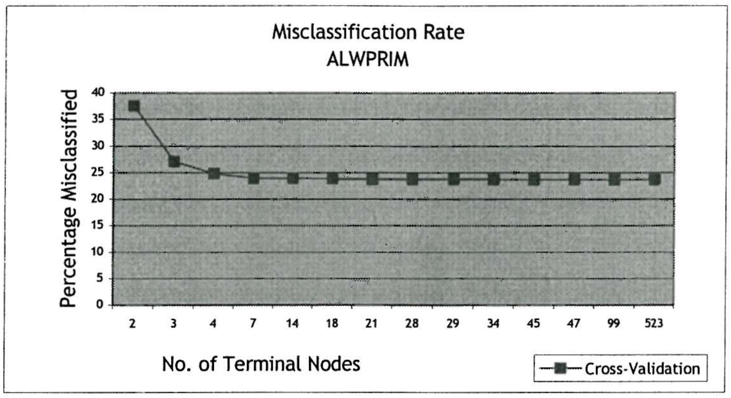


Figure 4.5.2.1 Misclassification Rate Plot

It can be seen that the misclassification rate clearly decreases until the tree has 7 terminal nodes and then it remains relatively constant. Similar patterns were observed in the misclassification rate figures of all other trees investigated in this analysis.

The three different sizes chosen were based on a compromise between the misclassification rate and the number of terminal nodes. This implied the use of misclassification rates, which were as small as possible with a manageable number of terminal nodes. Then, a "small" tree was defined as having around 7 terminal nodes, a "medium" tree with around 15 terminal nodes and a "large" tree with around 30 terminal nodes.
Since the misclassification rate is very stable after certain point as shown in *Figure 4.5.2.1*, large trees (larger that 30 terminal nodes) were not used due to time consuming in the imputation process.

CART software allows for an "optimal" tree to be built. As explained in *Chapter 2*, this is done by initially growing the largest possible tree and then pruning it back until a specified criterion is reached. This criterion is based on a compromise between the cost complexity of the tree (based on the number of terminal nodes) and its misclassification rate as shown in *Section 2.3.7*. Occasionally, the optimal tree could be the largest possible tree.
Whenever possible, the performance of the optimal tree was compared against the rest of those selected in order to evaluate differences between the optimal tree performance and other trees.

The next table present the results of the expected values of the point estimates and their variances for the four categories of the variable Ethnic when using different tree sizes. More information such as bias and standard deviations for this case can be seen in the *Appendix 2*.

Table 4.5.2.1

Expected values of the point estimates for the total and their variances depending on the number of terminal nodes used for the variable Ethnic

	Categories							
	1		2		3		4	
Nodes	$E(\hat{Y})$	$S(\hat{Y})$	$E(\hat{Y})$	$S(\hat{Y})$	$E(\hat{Y})$	$S(\hat{Y})$	$E(\hat{Y})$	$S(\hat{Y})$
2	142189	465	39724	403	7047	97	9796	151
3	142189	463	39725	365	7047	96	9796	141
4	142189	465	39724	366	7047	80	9796	133
10	142189	464	39724	366	7046	80	9797	136
13	142189	464	39724	366	7046	80	9797	136

In general, it can be seen from last table that neither the expected values of the point estimates nor their variances change very much when using different number of terminal nodes, however, we will see later in this chapter that using trees for forming the imputation classes does improve the imputations results. In this sense, we could say that there are not real differences between choosing a big or a small tree but only when taking into account computational and time resources.

4.5.3. Classifying the records for imputation

After the tree is grown using the observed cases for the auxiliary variable, the records with missing information in the response variable were "dropped down" the tree. This involved the identification of the terminal nodes (imputation classes) in which those records end up in the tree. The discovery of these terminal nodes also defines the pool of records from which imputed values will be obtained (pool of donors).

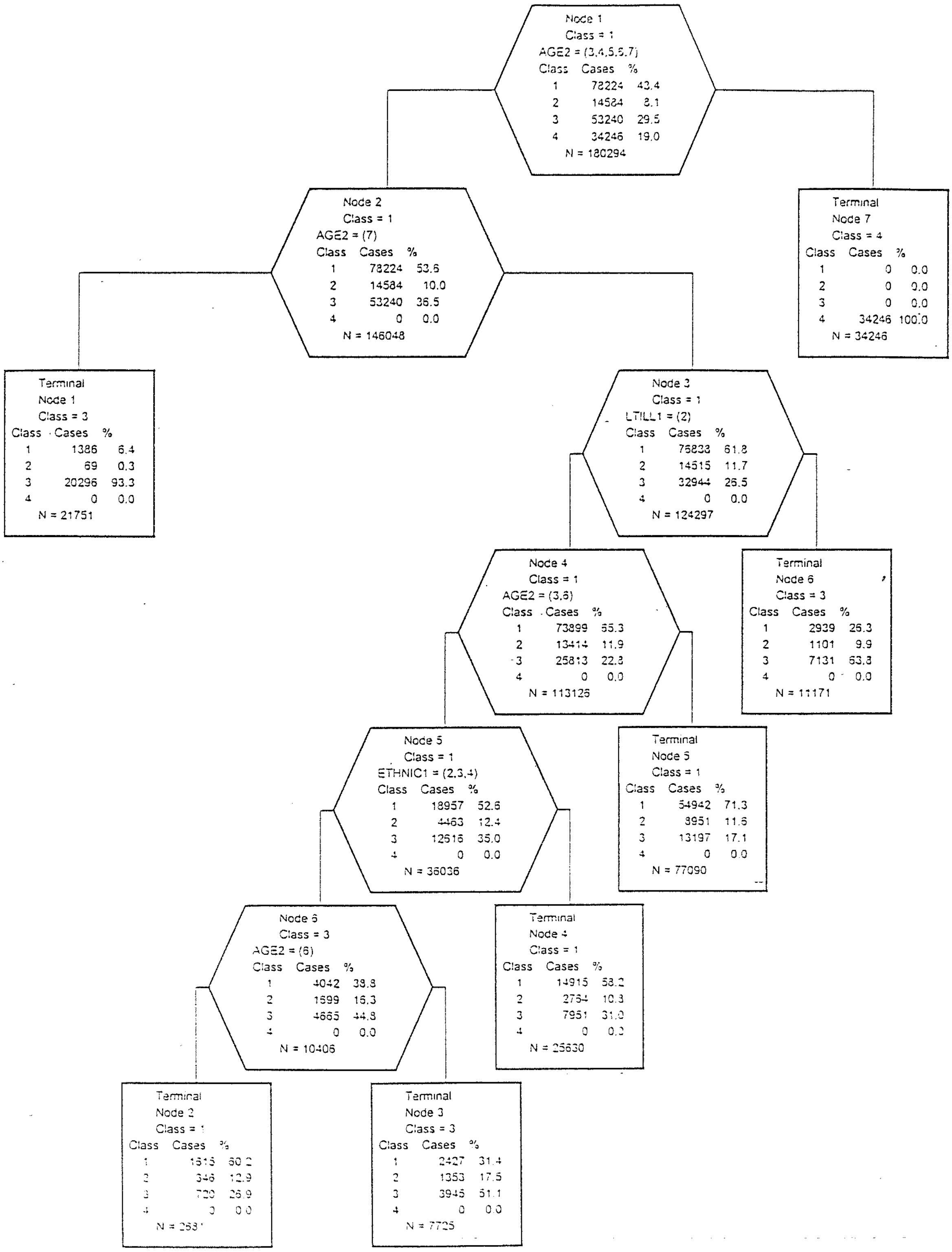
The process of dropping the records with missing information down the tree uses the set of rules (classification structure) that generate the tree. These rules categorise the records with missing information in the response variable depending on the values of the auxiliary variables. This requires each case to have sufficient information, in terms of the explanatory variables to allow for the classification to be done.

Then, the final result will be two different groups of records in each terminal node of the tree. One group is the pool of donors and the other group is the pool of recipients.

Figure 4.5.3.1 shows an example of a classification tree generated by CART

Figure 4.5.3.1

Example of a Classification Tree Generated by CART
(Variable: Primary Activity Last Week)



The tree shown in the last figure is a classification tree for the variable Primary Activity Last Week (ALWPRIM). There is a set of conditions associated with each terminal node, which define that node, and these are given by the classifiers as explained in *Section 2.3.1* in *Chapter 2*. For example, to reach terminal node 1, the record has to have values 3,4,5,6,or 7 in the variable AGE to be located in node 2 (instead of node 7) and then value 7 for the same variable AGE to be located in terminal node 1 (instead of node 3). After the whole tree is done, records with ALWPRIM missing but observed values for the rest of the variables can be dropped down the tree in order to be classified. The values of ALWPRIM missing can then be imputed from the range of values of AWLPRIM of those "observed" cases present in the specific terminal node where the records with missing values end up, using any of the imputation methods employed in the analysis.

After classifying all the records with missing information by the tree structure, the imputation was carried out independently at each terminal node. The imputation methods used for the analysis are explained in the next section.

4.6. IMPUTATION METHODS

Three different imputation methods were used in this research as described in *Chapter 3*. A brief reminder of those methods.

Frequency Distribution Method

As described in *Section 3.4* in *Chapter 3*, the probability distribution method can be applied in practice in two different ways, the way in which probability of having an specific class is given to the recipients depending on the probability distribution obtained in that terminal node, Probability Distribution imputation method itself; or the way in which the frequency distribution of the terminal node defines how many records will have an specific class assigned, Frequency Distribution imputation method.
This thesis uses the Frequency Distribution method instead of the Probability Distribution method due to its computational simplicity.

In the Frequency Distribution case the response distribution of the terminal node determines the values to be imputed. It means that the imputations will depend on the frequencies of the response variable in each terminal node.

Highest Probability Method (or Modal Imputation)

This method imputes the value that is "most likely" (i.e. has the highest probability) to all of the records with missing values in a specific terminal node as explained in *Section 3.4*.

Nearest Neighbour Method

As also described in *Section 3.4*, in this case, distances between the recipient and each possible donor within the node are calculated and the "nearest" donor defines the imputed value for that particular recipient. The nearest donor is then determined by the set of auxiliary variables. That is, the distance between the two records (recipient and possible donor) is calculated by the differences between their values for each of the auxiliary variables.

It is important to point out that a record can be used more than once as a donor. This means that if a record has the less distance to two different recipients it could be used as a donor to fill in the missing values for both of the recipients.
Moreover, when a recipient has the same distance to two different donors, one of the donors is selected randomly with equal probabilities.

All the imputation methods were applied to all of the tree sizes in order to obtain information about the relationship between the different imputation methods and the different tree sizes. The comparison is made in the results section.

4.7. EVALUATION OF IMPUTATION PERFORMANCE

4.7.1.Introduction

The evaluation should depend on the aims of the study and the information available for testing the results. This must be decided before the simulation is carried out and all the key aspects of the investigation must be taken into account.

In this work, we evaluate the imputation procedure as a whole, including the classification tree used. Different aspects are taken into account for carrying out the evaluation:

If a single variable is being imputed, as in this case, the evaluation of the performance of the imputation is based on:

✓ a comparison of marginal distributions (real and imputed values),

✓ a comparison of the individual values (real and imputed values for each single record)

✓ assessment of the properties of the estimator used

The results obtained from the simulations carried out can be analysed from different perspectives.

In general, different comparisons can be done depending on the area to be evaluated.

1. to assess the impact of using a classification tree for imputation, comparisons of the results of imputation using trees and not using trees can be done.
2. to evaluate the performance of the different imputation methods when using classification trees, comparisons between the results obtained using different imputation methods can be done.
3. to evaluate the properties of the estimators used in the analysis, bias and variances can be estimated.
4. in addition, if more details want to be given, comparisons can be made between the different categories of the variable being imputed.

In this thesis, the main aspect to be analysed are the differences in the imputation performance regarding the use of classification trees for forming the imputation classes.

In any case, the evaluation can be made by comparing marginal distributions and individual values before and after the imputation. For some authors, preservation of the distributions (both individual marginal distributions and joint distributions) and the preservations of individual values are the most important aspects to be evaluated when doing imputation. For others, bias and variances of the estimator used are more important. In this thesis, we try to cover as much as possible given the available resources.

The preservation of marginal distributions is essential to be assessed when the imputed data is going to be used for estimating aggregates or totals. In this case, preserving marginal distributions guarantee an accurate estimation of these aggregates, since individual values are not needed separately, such as in descriptive studies wherein only calculations of parameter as totals and proportions are needed. However, there are some cases where the micro data is required, as for example analytical analysis at individual levels, where it is important to maintain relationship between variables for the subjects. In these cases, where single cases can be needed, the preservation of individual values is a crucial aspect to be assessed when using imputation procedures.

In order to evaluate the three aspects mentioned at the beginning of this section, three different criteria were used. These are:

- ✓ Graphical comparison
- ✓ Test of agreement
- ✓ Bias and variance

All these include comparisons between real and imputed values.

The first two components of the evaluation are based on a comparison of the performance of the different combinations of classification trees and imputation methods. As can be seen, there are a very large number of possible combinations to analyse. For this reason, a description of the different methods used for evaluating these aspects will be given next and results of those, for all of the possible combinations, will be given in *Section 4.8*.

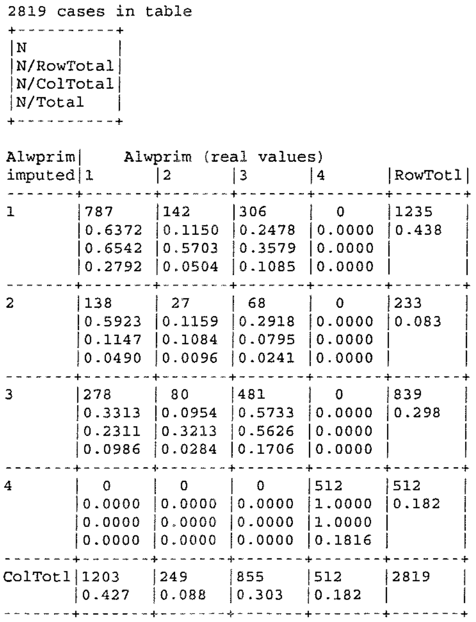
4.7.2. Graphical comparison

This evaluation consists of a comparison of the real and the imputed marginal distributions and individual values of the variables used in the analysis. The aim of this evaluation is to compare the real against the imputed marginal distribution in order to assess how the imputation preserves the original marginal distributions.

In order to carry out the comparisons mentioned above, two steps were followed. First, a cross-tabulation of the imputed values against the real values was produced. An example of this cross-tabulation is shown in the next table.

Table 4.7.2

Cross-Tabulation between real and imputed values for variable Primary Activity Last Week



```
2819 cases in table
+----------+
|N         |
|N/RowTotal|
|N/ColTotal|
|N/Total   |
+----------+

Alwprim|    Alwprim (real values)
imputed|1      |2      |3      |4      |RowTotl|
-------+-------+-------+-------+-------+-------+
1      |787    |142    |306    |  0    |1235   |
       |0.6372 |0.1150 |0.2478 |0.0000 |0.438  |
       |0.6542 |0.5703 |0.3579 |0.0000 |       |
       |0.2792 |0.0504 |0.1085 |0.0000 |       |
-------+-------+-------+-------+-------+-------+
2      |138    | 27    | 68    |  0    |233    |
       |0.5923 |0.1159 |0.2918 |0.0000 |0.083  |
       |0.1147 |0.1084 |0.0795 |0.0000 |       |
       |0.0490 |0.0096 |0.0241 |0.0000 |       |
-------+-------+-------+-------+-------+-------+
3      |278    | 80    |481    |  0    |839    |
       |0.3313 |0.0954 |0.5733 |0.0000 |0.298  |
       |0.2311 |0.3213 |0.5626 |0.0000 |       |
       |0.0986 |0.0284 |0.1706 |0.0000 |       |
-------+-------+-------+-------+-------+-------+
4      |  0    |  0    |  0    |512    |512    |
       |0.0000 |0.0000 |0.0000 |1.0000 |0.182  |
       |0.0000 |0.0000 |0.0000 |1.0000 |       |
       |0.0000 |0.0000 |0.0000 |0.1816 |       |
-------+-------+-------+-------+-------+-------+
ColTotl|1203   |249    |855    |512    |2819   |
       |0.427  |0.088  |0.303  |0.182  |       |
-------+-------+-------+-------+-------+-------+
```

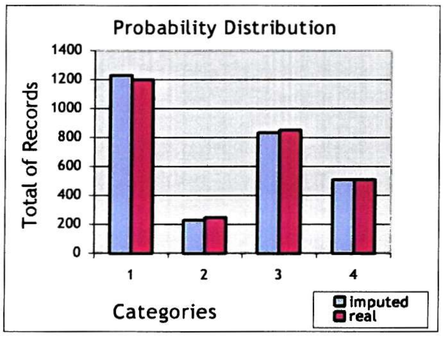
In this case, the imputation was made for the variable Primary Activity Last Week (records for which variable ALWPRIM is missing), and using a specific tree size and a specific method for imputation.

All the tables were produced using the software S-Plus.

Second, two different types of graphs were made. The first kind of graph is for comparing marginal distributions. An example of this graph is next.

Figure 4.7.2.1
Comparison of Marginal Distributions

Software: CART **Variable: ALWPRIM**
Marginals



Here, the blue columns represent the distribution of the imputed values for all of the categories of the variable used (ALWPRIM), and the red columns represent the distribution of the real values for the same categories of the variable. On the top of the figure, the name of the variable and the method used for imputation can be seen. Information about the size of the tree used is also included in the results section.

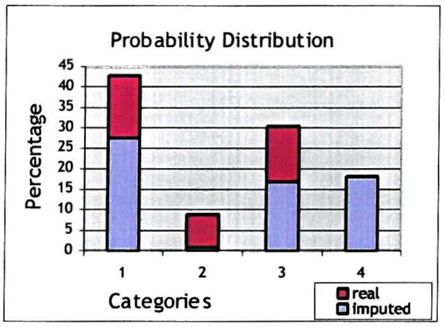
The second kind of graph is shown below and is used to compare how accurate the preservation of the individual values is. It compares each value of the variable before and after the imputation.

Figure 4.7.2.2

Comparison of Individual Values

Software: CART **Variable: ALPRIM**

Diagonals



In this figure, the blue part of the column represents the percentage of cases belong that category whose values were recovered by the imputation. On the other hand, the red part of the column represents the percentage of records that belong that category whose records were incorrectly imputed. It can be seen that all together represent the percentage of original records that belong to a specific category.

In this example, the percentage of records belonging to category 1 is 42.67%. After imputation, the percentage of records imputed correctly as a category 1 is 27.46% out of the total number of records in the database. This means, 64.35% of category 1 records were correctly imputed (27.46 out of 42.67) and the remaining 35.65% were imputed in any other category.

4.7.3. Test of Agreement

The aim of this evaluation is to determine whether or not marginal distributions, or even more individual values, are preserved after the imputation process is carried out.

Two different statistics were used for comparing marginal distributions and individual values.

The first comparison was between marginal distributions (imputed versus real) using a Wald Statistic proposed by Chambers (2000). This statistic tests how similar the two distributions are. Therefore, our null hypothesis is that both marginal distributions, imputed and real distributions, are equal versus the hypothesis that they are different.

The statistic has the form:

$$W = \left[\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^t\right]\left[\frac{1}{n^2}\sum_{i=1}^{n}(\hat{y}_i - y_i)(\hat{y}_i - y_i)^t\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)\right]$$

where $\hat{y}_i$ represents the imputed value and y_i the real values of $\mathbf{Y}$ for the $i\,th$ unit and n is the number of records in the cross-tabulation across the categories of the variable being imputed.

Under the hypothesis that the two marginal distributions (real and imputed) are equal, W should has an approximate chi-square distribution with p-1 degrees of freedom, where p is the order of the actual vs. imputed cross-tabulation mentioned before.

The second statistic proposed by Chambers (2000) is used for testing whether or not the individual values were maintained after the imputation was carried out. Therefore, our null hypothesis is the preservation of individual values of $\mathbf{Y}$ by the imputation. For simplification, the statistic is called "Diagonal" in this work and has the form:

$$z_D = \frac{D}{\sqrt{\hat{V}(D)}}$$

where D is the proportion of incorrectly imputed cases

$$D = 1 - n^{-1}\sum_{i=1}^{n} I(\hat{y}_i = y_i)$$

with estimated variance

$$\hat{V}(D) = \frac{1}{n} - \frac{1}{2n^2}\mathbf{1}'\left\{\sum_{i=1}^{n}\left[diag(\hat{y}_i - y_i)(\hat{y}_i - y_i)' - (\hat{y}_i - y_i)(\hat{y}_i - y_i)'\right]\right\}\mathbf{1}$$

Then, provided one cannot reject the hypothesis that the imputation method preserves the marginal distribution using the Wald statistic mentioned before, the preservation of individual values can be tested by using the confidence interval for D. In this case, $D - 2\sqrt{V(D)}$ should be less than zero in order to have some evidences that the individual values are preserved. In other words, if $z_D - 2 < 0$, then, the individual values can be said to be preserved.

Both statistics are described in detail in Chambers (2000). All the values for the both statistics and their p-values are shown in the results section.

An example of the table containing these results follows:

Table 4.7.3.1

Wald statistic for Primary Activity Last Week by imputation method and number of terminal nodes

Terminal Nodes	Frequency Distribution	Highest Probability	Nearest Neighbour
7	3.37	361.17	0.37
14	4.30	336.06	3.66
29	7.03	309.32	7.01

In this table, the columns represent the different imputation methods used and the rows the different tree sizes used. The numbers in the table are the values for the Wald statistic. Similar tables are presented for the p-values of the Wald statistic and for the values of the Diagonal statistics for all the variables, imputation methods and tree sizes.
This example was made using the same information used for the graphs shown before.

4.7.4. Bias and Variance

In order to study the properties of the estimators obtained by the imputation methods, biases and variances were estimated. These properties were studied in theory in *Chapter 3.* In this chapter, we assess their properties by simulation.

Simulation study

In order to obtain the bias and the variances of the estimators as well as estimates for the variance in some cases, a simulation was carried out. The simulation involved several steps, which are explained hereafter

1. ***Generation of the databases.*** First, 1000 databases were created. Each database is a simple random sample of 9241 units of the complete database (which contains 198756 records). These 1000 databases, which are called sample databases in this work, contain fully observed information. The size of the sample databases is such that it takes into account all the possible missing combinations used in the simulation study. The number of cases missing for each combination is shown in *Table 4.3.2.2* in *Section 4.3.2.*

2. ***Generation of the synthetic holes.*** The original percentages of missing information for the variables involved in the study were replicated on each of the sample databases using a simple random procedure without replacement. That is, for each combination missing, a simple random sample of records was selected without replacement from the

sample database in order to delete their values. The size of this selection depended on the number of cases missing for that specific combination. The variables for which the distributions are to be estimated in the univariate case are the same variables involved in the imputation process, these are Ethnic, Country of birth and Long-term illness. The total number of records missing in each variable is shown in *Table 4.3.2.2*. Thus, 7928 records were missing in each sample database for the univariate case.

3. ***Classifying the records for imputation***. In order to carry out the imputation, each record within each sample database was classified using the set of rules generated by the tree used. All the records (including those for which $\mathbf{Y}$ was missing) were classified into specific terminal nodes depending on the values of the auxiliary variables. Since the size of the tree does not seem to have an effect on the imputation results (see *Section 4.8.1* for the univariate results), only one tree size was used for the simulation.

Then, each sample was divided into the number of groups required depending on the size of the tree chosen for each specific variable to be imputed.

4. ***Imputation***. After each record was classified into its correspondent group, the imputation was carried out independently within those groups and then totalled in order to obtain the estimator required for the analysis. That is, 1000 estimates were obtained (each from each sample database)

5. ***Calculation of the biases and variances***. Once the 1000 estimates were obtained the bias, variance for the simulation and estimation of the variance (this was calculated only for two of the three imputation methods) were obtained. These calculations were based on the 1000 samples as follows

- $\widehat{Bias} = \dfrac{\sum_{s=1}^{1000} \hat{Y}^{(s)}}{1000} - Y$
- Variance is the variability obtained from the 1000 estimates for the total. It is important to point out that the true variance can be calculated by using the formulas obtained in *Chapter 3*. However, given that in some cases only the model-based approach was used, we do not have the probabilities for each category of the response variable given by the model. Therefore, the value taken as a true variance is the variability of the 1000 estimates obtained from the simulations.
- $\widehat{Variance} = \dfrac{\sum_{s=1}^{1000} \hat{V}^{(s)}}{1000}$. Each estimation of the variance (for each sample) was calculated by using the respective formula for the variance estimator obtained for each method.

Results and comments are shown in the results section.

It is important to point out that the approach undertaken for the generation of the sample databases (including the generation of the missing values) was essentially design-based. That is, the original database was used as a given population, and 1000 simple random samples were drawn from it (no model was used). This approach was used in order to simplify the simulation process.

4.7.5. Implementation of a Hot Deck procedure

As a basis for comparison a Hot Deck procedure was also implemented, since this is one of the most common imputation procedure employed in census data, as explained in *Section 1.7* in *Chapter 1*.

The procedure employed for the comparison is a Sequential Hot Deck (Little and Rubin, 1987; Kalton, 1983; Kalton and Kalsbeek, 1992; Madow, et al 1983) using two different cases, 1) two variables totally observed for creating the imputation cells and 2) three variables totally observed for creating the imputation cells. For example, in case 1, two fully observed variables were cross-classified in order to form the imputation groups in which imputation is carried out. Within each cell, the procedure imputes each missing record with the value of the previous record on the list. If the previous record has also missing information for the variable to be imputed, the previous of that one was used to impute both records missing.

The variables used for classifying the records before doing the imputation were AGE and SEX in the 2 classification variables case, and Age, Sex and Primary Activity Last Week in the case where 3 variables were used for classifying the records for the imputation.

There are some significant differences between the hot deck approach and the approach proposed in this thesis. First, in the case of hot deck, the classification is created by a simple cross-tabulation between the different categories of the variables involved, while in the procedure proposed in this thesis, the classification is created by classification tree which looks for the best way in which this classification can be done based on a learning sample and taking into account misclassification costs and complexity of the classification.

Second, the approach proposed in the thesis uses as many variable as it considers necessary for a more accurate classification, which can include not only two but more (even all of them) depending on the case, while the hot deck procedure uses only the variables that the analyst consider necessary based on experience and resources, which are normally not more than three.

Third, depending on the variables to be imputed, the proposed approach can create a new classification according to that specific variable, while hot deck procedure uses the same classification for most of the variables.

Four, the imputation procedure is different in both cases. In the hot deck case, the imputation is carried out in a sequential way as explained before, while the method proposed in this thesis uses three different imputation methods which are different from the sequential imputation.

4.8 RESULTS

4.8.1.Using trees

Table 4.8.1.1 shows the results of the values of the Wald statistic described in *Section 4.7.3* and their corresponding p-values.

Table 4.8.1.1
Wald statistic and p-values for the univariate case

			Wald Statistic			P-values		
Variable	Tree Size	d.f.	Freq. Dist.	High. Prob.	Near. Neig.	Freq. Dist.	High. Prob.	Near. Neig.
COB	6	4	2.90	168.25	5.67	0.57	0.00	0.22
	15	4	2.81	158.47	8.62	0.58	0.00	0.07
	18	4	3.38	163.32	3.97	0.49	0.00	0.40
	No Tree	4	3.04	369.00	2.76	0.55	0.00	0.59
ETHNIC	4	3	0.67	364.32	1.56	0.87	0.00	0.66
	10	3	0.77	379.97	0.13	0.85	0.00	0.98
	13	3	0.85	382.52	1.32	0.83	0.00	0.72
	No Tree	3	0.08	1006.00	2.88	0.99	0.00	0.40
LTILL	14	1	0.25	218.38	0.98	0.61	0.00	0.32
	21	1	0.29	202.24	0.82	0.58	0.00	0.36
	29	1	0.31	203.20	1.03	0.57	0.00	0.30
	No Tree	1	0.05	343.00	0.14	0.80	0.00	0.70

In this table, small values of the Wald statistics (or equivalently, big values for the p-value) suggest no evidence to reject the hypothesis that marginal distributions are maintained and vice versa. Since the degree of freedom for each variable varies depending on the number of categories (i.e. each variable has a different critical value for the test), we simplify the analysis by using p-values.

It can be noticed from this table that some variations can be found for the Wald Statistic, depending on the variable being imputed. However, given that all the p-values of the Wald statistic presented in *Table 4.8.1.1* for the Frequency Distribution and Nearest Neighbour, are bigger than 0.05, one can say that marginal distributions are preserved even when no trees are used.

In the case of Highest Probability methods, we can see that marginal distributions are never preserved, even when classification trees are used. Therefore, the first conclusion is that it seems to be no improvement in the preservation of marginal distribution with the inclusion of classification tree in the imputation process.

Table 4.8.1.2 contains the number of records imputed for each category for the variable ETHNIC using Highest Probability method.

Table 4.8.1.2
Number of records imputed by category and tree size for variable ETHNIC using Highest Probability method

Category	4 Term. nodes	10 Term. nodes	13 Term. nodes	No Tree
1	2961	2972	2974	3492
2	351	351	351	----
3	71	68	68	----
4	109	101	99	----

As mentioned before, in the case of Highest Probability, there is not preservation of marginal distribution in any of the different tree sizes, as observed in *Table 4.8.1.1*. However, it can be seen that some of the categories of the variable being imputed are represented in the imputed marginal distribution when using a tree, which does not happen in the case where classification trees are not used.

In this case, the use of the tree ensure the use of different categories when imputing, depending on the class assignment that define the terminal nodes class, while when trees are not used, the imputation will be made employing the category with highest probability in the whole database, which is just one as shown in *Tables 4.8.1.2*.

Similar patterns were found for the rest of the variables

Moreover, even when the use of a tree does not guarantee the preservation of the marginal distributions when using Highest Probability, an improvement in the value of the Wald statistic can be observed. There is a big gap between the values obtained when using trees and the value when trees are not used, as shown in *Table 4.8.1.1*. That means, there is a slight improvement in the preservation of marginal distributions when using trees. This improvement is because, even when there are some differences in the shape of the distribution, most of the categories of the imputed variable (sometimes all of them) are represented in the imputed distribution. In contrast, in the case where trees are not used, the imputed distribution is formed by only one single category, as explained in the last point. This can also be seen from the graphs in *Appendix 4* and *Appendix 5*.

Table 4.8.1.3 contains the values of the Diagonal Statistic for the variables Country of Birth, Ethnic and Long Term Illness for the combination between different imputation methods and different tree sizes.

Table 4.8.1.3

Diagonal Statistic values for the univariate case

Variable	Tree Size	Freq. Dist.	High. Prob.	Near. Neig.
COB	6	12.22	6.72	11.09
	15	11.84	6.75	10.52
	18	11.56	6.78	11.22
	No Tree	19.95	10.68	11.60
ETHNIC	4	24.22	13.37	19.50
	10	23.63	13.33	20.32
	13	23.53	13.33	19.17
	No Tree	34.66	20.17	19.67
LTILL	14	9.56	6.41	9.07
	21	9.85	6.20	8.85
	29	9.22	6.22	8.58
	No Tree	12.72	6.81	9.13

Provided one cannot reject the hypothesis that the imputation method preserves the marginal distribution using the Wald statistic as pointed out before, the preservation of individual values can be tested by using the confidence interval for D (proportion of incorrectly imputed cases) as explained in *Section 4.7.3*. In this case, $D-2\sqrt{V(D)}$ must be less than zero in order to have some evidences that the individual values are preserved. In other words, if $z_D - 2 < 0$, then, the individual values can be said to be preserved, with $z_D = \frac{D}{\sqrt{\hat{V}(D)}}$.

Then, if a confidence interval is calculated using the information provided by *Table 4.8.1.3*, we can observed that $z_D - 2$ is closer to zero when trees are used than in the case of not using trees. This improvement can be observed from the point of view of percentage of records correctly imputed and it will be explained later.

In the case of Nearest Neighbour method, the values of z_D are very similar in both cases, when using or not trees. Then, it cannot be said the method performs better when using trees than when trees are not used.

It can also be noted from *Table 4.8.1.3* that there is a difference between the values of the statistic when using trees and the values when trees are not used for the first two methods Frequency Distribution and Highest Probability imputation. Then, in these cases, even when the diagonal statistics results indicate that the individual values are not preserved, the values of this statistic in the case of using trees are lower than the values when trees are not used.

Therefore, another general conclusion for the univariate case is that the use of the tree improves the performance of the imputation results in terms of preservation of individual values depending on the method used.

Tables 4.8.1.4 (as well as *Appendix 7*) present the "improvement" for the different combinations between tree sizes and imputation methods for the variables used in the univariate case.

Table 4.8.1.4

Improvement by variable, tree size and imputation method for the univariate case

Variable	Tree Size	Freq. Distrb.	High. Prob.	Near. Neigh.
COB	6	21.11	10.48	1.28
	15	22.28	10.40	2.74
	18	23.12	10.31	0.94
	No Tree	0.00	0.00	0.00
ETHNIC	4	18.69	12.06	0.27
	10	19.86	12.14	-1.07
	13	20.07	12.14	0.83
	No Tree	0.00	0.00	0.00
LTILL	4	6.03	0.75	0.12
	21	5.46	1.14	0.53
	29	6.69	1.10	1.03
	No Tree	0.00	0.00	0.00

This measure of improvement is calculated based on the percentage of records correctly imputed when trees are not used and their differences with the percentage of records correctly imputed when trees are used. For example, the percentage of records correctly imputed for Country of birth when using a tree with 6 terminal nodes and Frequency

Distribution as imputation method is 73.48%, and the same percentage but in the case when trees are not used is 60.67%. Then, the improvement when using trees with respect to the case where trees are not used is 21.12%, which correspond to (73.48 - 60.67) / 60.67. Therefore, it can be said that there is an improvement in the performance of the imputation method when using a classification tree in about 21% compared to the case where trees are not used.

If a comparison between the results from the case where trees are used and the case where trees are not used is made, we will notice that there is always an improvement in terms of records correctly imputed when using trees for the Frequency Distribution and almost always for the Highest Probability method. This improvement can reach more than 20% in some cases for the univariate case.

It is clear that the highest improvement is always for Frequency Distribution method, followed by Highest Probability and Nearest Neighbour as the last one with almost no improvement.

Therefore, we can say that even when the values of the diagonal statistic in the case of using tree reveal that individual values are not preserved, we can confirm that there is an improvement on the percentage of records correctly imputed when using trees.

4.8.2.Comparing Tree-Sizes

Table 4.8.1.1 and *Table 4.8.1.3* show that there are some differences between the values for both Wald and Diagonal statistics when the tree size is changed. However, since the p-values for the Wald statistic are over 0.05, there is not enough evidence to reject the hypothesis that the individual marginal distributions are preserved in any of the cases. Similar conclusion can be drawn in the case of the Diagonal Statistic, where all the values, even when there are some differences, are big enough to confirm that individual values are not maintained.

Therefore, the main conclusion about using different sizes for the tree is that increasing the size does not necessarily improve the imputation performance. The results obtained from the analysis show that the changes on the Wald statistic and the Diagonal statistic are not big enough to alter the conclusion that the imputation performance is not affected by the size of the tree.

Additionally, the changes on both statistics do not follow similar pattern for all of the cases. Sometimes the best results are obtained from the smallest trees and sometimes from the biggest trees or even from the medium size trees.

Then, an important conclusion is that using complex trees does not necessarily lead to better imputation results.

Moreover, *Table 4.8.1.1* and *Table 4.8.1.3* include the Wald statistic and the Diagonal statistic for the variable COB. In this case, three different tree sizes were used, in which the biggest is also the optimal tree given by CART. It is clear from the results that even when this optimal tree is used, there are not considerable differences in the results when comparing both marginal distributions and individual values.

Therefore, we can say that the use of the optimal tree given by CART does not make major improvement in the performance of the imputation. The optimal tree given by CART is meant to be optimal in terms of complexity and misclassification rate. In this sense, the use of the optimal tree could be expected to give the best performance, however, it can be observed from the results that this hypothesis is not necessarily correct.

Table 4.8.2.1 contains the percentage of missing data for each variable, the percentage of records incorrectly imputed for the different imputation methods as well as the misclassification rate for each specific tree.

Table 4.8.2.1

Percentage of missing data, Percentage of records incorrectly imputed and misclassification rate by variable, imputation method and tree size for the univariate case.

Variable	% Miss. data	Tree Size	Freq. Dist.	High. Prob.	Near. Neig.	Miscl. Rate
Cob	7.26	6	26.52	15.63	24.40	14.99
		15	25.81	15.69	23.31	14.60
		18	25.30	15.75	24.66	14.57
		No tree	39.33	23.63	25.36	------
Ethnic	16.24	4	33.44	20.21	28	18.87
		10	32.78	20.16	28.98	18.82
		13	32.67	20.16	27.6	18.81
		No tree	43.92	28.80	28.20	------
Ltill	13.37	4	16.31	11.26	15.54	11.05
		21	16.76	10.92	15.20	10.99
		29	15.79	10.95	14.78	10.97
		No tree	21.07	11.93	15.65	------

There seems to be a relationship between the misclassification rate and the percentage of records incorrectly imputed for each variable. It can be seen from this table (and from the set of graphs in *Appendix 8*) that the percentage of records incorrectly imputed increases when the misclassification rates increases and even more when trees are not used. Additionally, the percentage of records incorrectly imputed look stable as well as the misclassification rate within each variable for each imputation method.

Also, we can see that these results are not related to the size of the tree, that is, percentage of records incorrectly imputed and misclassification rate look very stable across the different tree sizes.

4.8.3.Comparing Imputation Methods

Table 4.8.1.1 and *Table 4.8.1.3* include the p-value for the Wald statistic and the Diagonal statistic for the case of the variable Country of birth. The values for this value illustrate how the individual marginal distribution for this variable is maintained and how individual values are not preserved.

However, on the other hand, *Table 4.8.3.1* shows the total of records correctly imputed (including values for variable COB) depending on the size of the tree when using Frequency Distribution method for imputation. It is clear that the use of the tree increases this numbers with respect to the case where trees are not used. Similar patterns can be observed for ETHNIC and LTILL even when the differences are smaller in the last case.

Table 4.8.3.1

Total of records correctly imputed by variable, tree size and imputation method for the univariate case

Variable	Tree Size	Freq. Distrb.	High. Prob.	Near. Neigh.	
COB	6	1147	1317	1180	
	15	1158	1316	1197	
	18	1166	1315	1176	
	No Tree	947	1192	1165	(total of cases: 1561)
ETHNIC	4	2324	2786	2514	
	10	2347	2788	2480	
	13	2351	2788	2528	
	No Tree	1958	2486	2507	(total of cases: 3492)
LTILL	4	2406	2551	2428	
	21	2393	2561	2438	
	29	2421	2560	2450	
	No Tree	2269	2532	2425	(total of cases: 2875)

Table 4.8.3.2 shows these results in term of percentages. *Appendix 6* shows a graphical representation of this table by variable.

Table 4.8.3.2

Percentage of records correctly imputed by variable, tree size and imputation method for the univariate case

Variable	Tree Size	Freq. Distrb.	High. Prob.	Nera. Neigh.
COB	6	73.47	84.36	75.59
	15	74.18	84.30	76.68
	18	74.69	84.24	75.33
	No Tree	60.66	76.36	74.63
ETHNIC	4	66.55	79.78	71.99
	10	67.21	79.83	71.01
	13	67.32	79.83	72.39
	No Tree	56.07	71.19	71.79
LTILL	4	83.68	88.73	84.45
	21	83.23	89.07	84.80
	29	84.20	89.04	85.21
	No Tree	78.92	88.06	84.34

Therefore, in the case of Frequency Distribution, there is always an improvement when using tree. This improvement is not evident when comparing marginal distributions but it can be observed when comparing individual values.

Next example contains the values for the Wald statistic (p-value) and the Diagonal statistic for the variable Ethnic.

Tables 4.8.3.3

Wald Statistic, p-value and diagonal statistic for the variable Ethnic

Wald Statistic

Tree Size	Prob. Distrb.	High. Prob.	Near. Neigh.
4	0.67	364.32	1.56
10	0.77	379.97	0.13
13	0.85	382.52	1.32
No Tree	0.08	1006.00	2.88

Diagonal Statistic

Tree Size	Prob. Distrb.	High. Prob.	Near. Neigh.
4	24.22	13.37	19.50
10	23.63	13.33	20.32
13	23.53	13.33	19.17
No Tree	34.66	20.17	19.67

Wald Statistics P-value

Tree Size	Prob. Distrb.	High. Prob.	Near. Neigh.
4	0.87	0.00	0.66
10	0.85	0.00	0.98
13	0.83	0.00	0.72
No Tree	0.99	0.00	0.40

These tables show that there are not big changes neither when the size of the tree is altered nor when trees are not used when using Nearest Neighbour method. The behaviour of the Nearest Neighbour method remains relatively constant in this sense.

It is important to point out that, in general, the use of trees does not make any improvement in the results when using Nearest Neighbour, probably because the nearest neighbour donor will be found either using or not classification. We can see that the results remain the same when comparing both marginal distributions and individual values. Additionally, the percentage of records correctly imputed remains fairly stable when using Nearest Neighbour as showed in *Table 4.8.3.2*. The use of the tree will probably improve the time consumed in the imputation process given that donors will be only sought in the corresponding terminal node.

More examples of this point can be found in *Table 4.8.1.1* and *Table 4.8.1.3* in *Section 4.8.1.*

It can be seen from *Table 4.8.3.2* that the best method in preserving individual values is the Highest Probability with up to almost 90% of the cases correctly imputed in some situations. The percentage of records correctly imputed with this method depends, in a way, on the shape of the distribution when using trees and of course on the accuracy of the classification tree.

Thus, in general, the best methods for preserving marginal distributions are Frequency Distribution and Nearest Neighbour. These two methods perform very well even when trees are not used, which is not the case of the Highest Probability method. Example can be found in *Appendix 4*, which are graphical representations of the Wald statistic from *Table 4.8.1.1*. However, in terms of preservation of individual values, Highest Probability seems to be the best performing method.

4.8.4.Comparing Categories

Tables 4.8.4.1 contain the percentage of records incorrectly imputed by imputation methods tree sizes and categories, as well as the misclassification rates obtained from the different tree sizes by categories of the different target variables.

Tables 4.8.4.1

Misclassification rates by tree sizes and categories and percentage of records incorrectly imputed by imputation method, tree size and categories for the univariate case

Table A
Variable: COUNTRY OF BIRTH

					Percentage of Records Incorrectly Imputed											
		Misclass. Rate			Freq. Dist.				High. Porb.				Near. Neigh.			
Cat.	Records	8 TN	15 TN	27 TN	8 TN	15 TN	27 TN	N TRE	8 TN	15 TN	27 TN	N TRE	8 TN	15 TN	27 TN	N TRE
1	137958	1.47	1.68	1.61	14.18	13.59	13.26	22.48	1.34	1.59	1.51	0.00	13.59	11.24	13.09	14.35
2	11608	100.00	100.00	100.00	94.21	90.91	90.91	95.04	100.00	100.00	100.00	100.00	85.95	90.91	91.74	92.56
3	3648	57.15	43.75	43.01	72.50	70.00	62.50	97.50	57.50	42.50	40.00	100.00	70.00	65.00	52.50	60.00
4	18735	36.94	37.59	37.97	43.11	43.11	41.92	91.62	34.73	38.32	40.12	100.00	35.33	37.13	40.72	34.13
5	5287	74.48	62.74	62.74	73.17	82.93	78.05	95.12	63.41	58.54	58.54	100.00	68.29	78.05	70.73	78.05

Table B
Variable: ETHNIC

					Percentage of Records Incorrectly Imputed											
		Misclass. Rate			Freq. Dist.				High. Porb.				Near. Neigh.			
Cat.	Records	4 TN	10 TN	13 TN	4 TN	10 TN	13 TN	N TRE	4 TN	10 TN	13 TN	N TRE	4 TN	10 TN	13 TN	N TRE
1	126733	2.68	2.47	2.40	19.95	19.55	19.47	28.20	3.14	2.90	2.86	0.00	16.25	17.06	15.33	16.37
2	35416	58.03	58.03	58.03	63.03	60.91	60.76	77.76	60.34	60.34	60.34	100.00	51.56	53.26	51.56	52.27
3	6286	58.10	58.38	58.38	69.84	71.43	69.05	94.44	63.49	63.49	63.49	100.00	61.90	65.87	69.84	65.87
4	8801	66.41	68.34	69.19	79.89	79.89	81.03	94.83	70.11	72.41	72.99	100.00	75.86	74.14	75.29	72.41

Table C
Variable: LONG TERM ILLNESS

					Percentage of Records Incorrectly Imputed											
		Misclass. Rate			Freq. Dist.				High. Porb.				Near. Neigh.			
Cat.	Records	14 TN	21 TN	29 TN	14 TN	21 TN	29 TN	N TRE	14 TN	21 TN	29 TN	N TRE	14 TN	21 TN	29 TN	N TRE
1	20753	86.99	84.16	84.54	69.97	72.01	67.93	89.21	86.01	82.51	82.80	100.00	68.22	66.47	65.01	66.76
2	156483	0.98	1.30	1.22	9.04	9.28	8.73	11.85	1.15	1.22	1.22	0.00	8.41	8.25	7.98	8.73

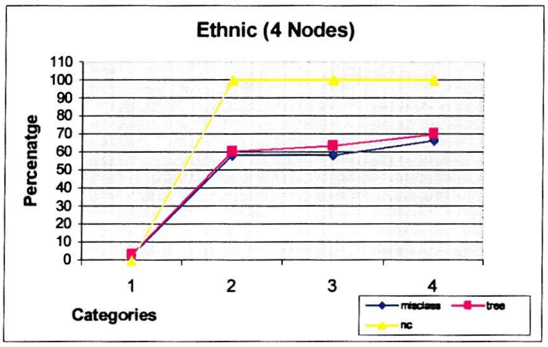
It can be noticed from these tables that there seems to be a relationship between the misclassification rates obtained from the tree and the percentage of records incorrectly imputed by categories. Both the misclassification rate and the percentage of records incorrectly imputed by categories tend to follow similar patterns most of the time. It can also be observed that this relationship is not necessarily the same for the case where trees are not used.

Another interesting finding obtained from this table is that in the case when trees are not used, the percentage of records incorrectly imputed by categories is usually higher (or at least equal) than the percentage of records incorrectly imputed when trees are used for the Frequency distribution and Highest Probability methods.
Moreover, depending on the imputation method used, the percentage of records incorrectly imputed obtained from the case where trees are not used can be near to 100% for most of the categories as is the case of Highest Probability method where only one category is used for imputation. This corroborates the statement made previously that the use of trees improves the performance of the imputation results depending on the method used.

In the case of Nearest Neighbour method, all the information, percentage of records incorrectly imputed using trees, percentage of records incorrectly imputed when trees are not used and misclassification rate, have more similar results across categories than the rest of the methods. As said before, this implies that there is not an impact on the imputation results when Nearest Neighbour method together with classification trees is used for imputation.

A set of graphs obtained from *Tables 4.8.4.1* can be found in the *Appendix 9.* These graphs show the percentage of records incorrectly imputed using and not using trees and the misclassification rate by categories for the different imputation methods and different tree sizes.
Figure 4.8.4.1 is an example of the set of graph presented in *Appendix 9*. This figure shows the percentage of records incorrectly imputed using and not using trees and the misclassification rate by categories for the variable Ethnic, using the Highest Probability method and a tree with 4 terminal nodes.

Figure 4.8.4.1
Misclassification rate versus percentage of records incorrectly imputed by categories for variable Ethnic (*Appendix 9*)



It can be seen from the figure how the lines for the percentage of records incorrectly imputed obtained using trees (red line) and the misclassification rates for the same categories (blue line) follow the same pattern. Alternatively, the line representing the percentage of records incorrectly imputed in the case where trees are not used (yellow line) is different from the two lines mentioned before.

There seems to be a relationship between the misclassification rate (blue lines) and the percentage of records incorrectly imputed (red line) for each category when trees are used. It means that both the misclassification rate and the percentage of records incorrectly imputed by categories tend to have the same values or at least follow similar patterns most of the time.

It can also be observed that this relationship with the percentage of records incorrectly imputed (red line in the graph) is not necessarily the same for the case where trees are not used (yellow lines in the graph).

This is an important finding from the point of view of accuracy. It could be predicted from the tree, by using the misclassification rate by categories, which categories of the variable being imputed will be more accurate than others after the imputation is done.

4.8.5. Bias and Variance Results

Given the results obtained from the previous analysis that the size of tree is not directly related to the imputations results, the simulations for the bias and variances were carried out using only one tree size. The size chosen for this analysis was the medium tree size (about 15 nodes) which represent a reasonable number of groups to work with, as the time consuming for the variance simulations is in fact very long, specially in the case of the Nearest Neighbour imputation.

Each section presents a set of summary tables, more detailed information can be seen in *Appendix 3.*

4.8.5.1. Bias

Tables 4.8.5.1.1 contain the bias results obtained from the simulations described in *Section 4.7.4* for all the variables and imputation methods used in the univariate case.

Tables 4.8.5.1.1
Biases estimation for the univariate case

Table A
Variable: Country of birth

	Frequency Distribution					Highest Probability					Nearest Neighbour				
	Categories					Categories					Categories				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
$E(\hat{Y})$	154760	13046	4060	21015	154760	154915	12944	4058	20975	5864	154760	13046	4060	21015	5874
$E(\hat{Y})-Y$	0.36	-0.36	-0.40	-0.47	0.36	155.00	-102.00	-2.00	-40.00	-11.00	0.21	0.36	-0.15	0.09	-0.50
$((E(\hat{Y})-Y)/Y)\cdot 100$	0.000	-0.002	-0.009	-0.002	0.014	0.100	-0.781	-0.049	-0.190	-0.187	0.000	0.002	-0.003	0.000	-0.008

	Hot Deck (clas.: 2 var.)					Hot Deck (clas.: 3 var.)				
	Categories					Categories				
	1	2	3	4	5	1	2	3	4	5
$E(\hat{Y})$	154760	13046	4060	21015	5875	154759	13046	4060	21015	5875
$E(\hat{Y})-Y$	0.00	0.00	0.00	0.00	0.00	-1.00	0.00	0.00	0.00	0.00
$((E(\hat{Y})-Y)/Y)\cdot 100$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table B
Variable: Ethnic

	Frequency Distribution				Highest Probability				Nearest Neighbour			
	Categories				Categories				Categories			
	1	2	3	4	1	2	3	4	1	2	3	4
$E(\hat{Y})$	142189	39724	7046	9797	142653	39496	6993	9714	142190	39724	7047	9795
$E(\hat{Y})-Y$	-0.34	0.03	-1.07	1.37	464.00	-328.00	-54.00	-82.00	0.64	-0.35	0.30	-0.58
$((E(\hat{Y})-Y)/Y)\cdot 100$	0.000	0.000	-0.015	0.013	0.326	-0.825	-0.766	-0.837	0.000	0.000	0.004	-0.005

	Hot Deck (clas.: 2 var.)				Hot Deck (clas.: 3 var.)			
	Categories				Categories			
	1	2	3	4	1	2	3	4
$E(\hat{Y})$	142189	39724	7046	9797	142190	39724	7046	9796
$E(\hat{Y})-Y$	0.00	0.00	-1.00	1.00	1.00	0.00	-1.00	0.00
$((E(\hat{Y})-Y)/Y)\cdot 100$	0.000	0.000	-0.014	0.010	0.000	0.000	-0.014	0.000

Table C
Variable: Long term illness

	Frequency Distribution		Highest Probability		Nearest Neighbour	
	Categories		Categories		Categories	
	1	2	1	2	1	2
$E(\hat{Y})$	23217	175539	22950	175806	23217	175539
$E(\hat{Y})-Y$	0.16	-0.16	-267.00	267.00	0.10	-0.10
$((E(\hat{Y})-Y)/Y)*100$	0.000	0.000	-1.150	0.152	0.000	0.000

	Hot Deck (clas.: 2 var.)		Hot Deck (clas.: 3 var.)	
	Categories		Categories	
	1	2	1	2
$E(\hat{Y})$	23217	175539	23217	175539
$E(\hat{Y})-Y$	0.00	0.00	0.00	0.00
$((E(\hat{Y})-Y)/Y)*100$	0.000	0.000	0.000	0.000

It can be seen from these tables that the Frequency Distribution and Nearest Neighbour methods lead to essentially unbiased estimates. That is, the difference between the real total and its expected value obtained from the simulations are basically zero. However, that difference is increased when the Highest Probability imputation method is used, as expected from the theoretical results. It can be noticed that even when some of the bias results for this method are about 400, the relative bias results show that these values are not big with respect to the real total.

Additionally, *Tables 4.8.5.1.1* show that the bias is positive for the major category of the variable and negative for the rest in the case of Highest Probability method. That means, the category containing more information is always overestimated and the rest are always underestimated. This occurs because the method imputes all the records with missing information using the major frequency in the node, which is the category containing most records.

In terms of the Hot Deck imputation, we can see that the estimator of the total is unbiased as well as in the case of Frequency Distribution and Nearest Neighbour imputation methods. Therefore, there are not major differences in terms of using any of the imputation methods (including Hot Deck) for estimating the total of cases in each category of the variables used for the analysis, except for the case of Highest Probability method.

4.8.5.2 Variance

Tables 4.8.5.2.1 contains the information related to the variances and variances estimation (in some cases) for the different variables when using different imputation methods.

Tables 4.8.5.2.1
Variances and Expected Variances Estimators for the univariate case

Table A
Variable: Country of birth

	Frequency Distribution					Highest Probability					Nearest Neighbour				
	Categories					Categories					Categories				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
$E(\hat{V})$	167.97	94.63	20.34	69.14	37.11	-	-	-	-	-	332.16	188.50	40.48	135.87	332.16
$V(\hat{Y})$	160.80	94.27	19.25	64.30	38.11	171.00	94.00	24.00	73.00	171.00	319.22	181.47	35.70	122.06	319.22
$MSE(\hat{Y})$	160.93	94.40	19.41	64.52	38.87	24196.0	10498.0	28.00	1673.00	168.00	319.26	181.60	35.72	122.07	66.05

	Hot Deck (clas.: 2 var.)					Hot Deck (clas.: 3 var.)				
	Categories					Categories				
	1	2	3	4	5	1	2	3	4	5
$E(\hat{V})$	-	-	-	-	-	-	-	-	-	-
$V(\hat{Y})$	441.00	169.00	49.00	256.00	81.00	441.00	169.00	49.00	256.00	81.00
$MSE(\hat{Y})$	441.00	169.00	49.00	256.00	81.00	442.00	169.00	49.00	256.00	81.00

Table B
Variable: Ethnic

	Frequency Distribution				Highest Probability				Nearest Neighbour			
	Categories				Categories				Categories			
	1	2	3	4	1	2	3	4	1	2	3	4
$E(\hat{V})$	489.58	419.79	85.17	137.27	-	-	-	-	970.78	831.78	168.75	271.95
$V(\hat{Y})$	463.76	365.97	79.83	136.23	498.00	398.00	86.00	150.00	797.95	671.59	157.61	268.23
$MSE(\hat{Y})$	463.88	365.97	80.97	138.11	215794	107982	3002.00	6874.00	798.36	671.71	157.70	268.57

	Hot Deck (clas.: 2 var.)				Hot Deck (clas.: 3 var.			
	Categories				Categories			
	1	2	3	4	1	2	3	4
$E(\hat{V})$	-	-	-	-	-	-	-	-
$V(\hat{Y})$	961.00	729.00	225.00	289.00	1024.00	784.00	196.00	289.00
$MSE(\hat{Y})$	961.00	729.00	226.00	290.00	1025.00	784.00	197.00	289.00

Table C
Variable: Long term illness

	Frequency Distribution		Highest Probability		Nearest Neighbour	
	Categories		Categories		Categories	
	1	2	1	2	1	2
$E(\hat{V})$	232.55	232.55	-	-	461.94	461.94
$V(\hat{Y})$	226.69	226.69	291.00	291.00	451.90	451.90
$MSE(\hat{Y})$	226.72	226.72	71580.0	71580.0	451.91	451.91

	Hot Deck (clas.: 2 var.)		Hot Deck (clas.: 3 var.)	
	Categories		Categories	
	1	2	1	2
$E(\hat{V})$	-	-	-	-
$V(\hat{Y})$	484.00	484.00	400.00	400.00
$MSE(\hat{Y})$	484.00	484.00	400.00	400.00

It can be noticed from last tables that there are not big differences in the values of the variances between Frequency Distribution and Highest Probability methods; however, there are more notable differences between these two methods and the Nearest Neighbour method. The Nearest Neighbour method produces bigger variances than the other two methods.

In term of Hot Deck imputation we can see that there is a difference between the variance obtained by this method and variance obtained by any of the rest of the imputation methods employed in this analysis. We can see that the variance obtained by Hot Deck is always higher than the variance obtained by any of the methods, even for Nearest Neighbour method, which produces the biggest variances among all of the three imputation methods used in the proposed approach.

Even when a third variable was included in the classification prior to the imputation when using Hot Deck, the results were still very similar.

However, in terms of mean square errors, Hot Deck provides smaller MSE than Highest probability method given the bias of the latest. Therefore, it can be said that any of the imputation methods proposed in this thesis perform better than the normal sequential Hot Deck imputation method except for the case of Highest Probability imputation method.

4.8.5.3. Variance Estimation

In terms of the variability, we can see from *Tables 4.8.5.2.1* that the estimator of the variance in the case of Frequency Distribution is basically unbiased as demonstrated in the theory in *Chapter 3*. Some small bias can be found depending on the variable used, however, these bias are very small compared with the size of the point estimator. The variable with less bias for the variance estimation is LTILL followed by COB and ETHNIC respectively. The same pattern can be found for the Nearest Neighbour imputation method.

On the other hand, even when we have proof that the estimator of the variance is unbiased in theory in the case of Nearest Neighbour imputation method, we can notice some differences between the real value and the expected values of the estimator over the 1000 simulations.
It can be seen from *Tables 4.8.5.2.1* that these bias can reach up to over 20% in few cases. However, these biases are present only in few cases. It is important to point out that these differences are considered high for estimating the variance but they are low in relation to the size of the point estimator.

Another important issue about the estimator of the variance in the case of Nearest Neighbour, and also in any small difference found for the Frequency Distribution method, is the fact that the variance is always overestimated.

Finally, we can confirm that in terms of variance estimation that, given the results of the simulations carried out, Frequency Distribution is in general the best performing imputation methods.

4.8.5.4. Coverage

Tables 4.8.5.4.1 show the results for the coverage for the different variables and imputation methods given by the simulations.

Tables 4.8.5.4.1
Coverage for the univariate case

Table A
Variable: Country of birth

	Frequency Distribution					Highest Probability					Nearest Neighbour				
	Categories					Categories					Categories				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Coverage	95.8	95.50	95.10	96.60	95.10	0.00	0.00	92.20	0.10	66.20	96.00	95.00	95.20	95.50	96.10

Table B
Variable: Ethnic

	Frequency Distribution				Highest Probability				Nearest Neighbour			
	Categories				Categories				Categories			
	1	2	3	4	1	2	3	4	1	2	3	4
Coverage	94.90	95.60	94.80	95.00	0.00	0.00	0.00	0.00	95.30	94.40	96.40	95.40

Table C
Variable: Long term illness

	Frequency Distribution		Highest Probability		Nearest Neighbour	
	Categories		Categories		Categories	
	1	2	1	2	1	2
Coverage	95.40	95.40	0.00	0.00	95.30	95.30

Tables 4.8.5.4.1 show how the coverage, that is the proportion of intervals for the estimator that include the parameter, is over 94% all the time in both Frequency Distribution and Nearest Neighbour. However, in the case of Highest Probability methods, only two cases are not 0% coverage. It can be seen from *Tables 4.8.5.1.1* that these two cases contain very few units (about 3% of the population).

It is important to point out that the confidence intervals in the case of Highest Probability were estimated using the values of the variance instead of the variance estimates given that the latest were not obtained in the simulations carried out in this thesis.

A reason for this coverage problem in the Highest Probability case is that even when the variance seems to be as big as the variance in the case of Frequency Distribution imputation method (method which has over 94% coverage), the bias is big enough to produce these results. In this case, we can see that the size of the bias is as big as the size of the variance (or sometimes bigger), which does not occur in the Frequency Distribution case, making the coverage very poor.

Additionally, there is not visible pattern in the case of coverage neither by imputation methods nor by variables. That is, not all the small categories have the less coverage or vice-versa.

In conclusion, there are some general findings we can summarise in terms of the analysis for the univariate case.

In general, the use of classification trees does improve the performance of the imputation. As seen in the results this improvement cannot be seen from the point of view of the maintenance of marginal distributions in most of the cases but from the point of view of percentage of records correctly imputed.

Even when there are differences in the results when using or not trees (i.e. when trees improve the performance of the imputation results), the use of different tree sizes does not have a major impact on those results. Moreover, the use of the optimal tree given by CART does not make much difference on the results.

Frequency Distribution and Nearest Neighbour methods preserve marginal distributions while Highest Probability does not. However, Highest Probability is the best performing imputation method.
In the case of Nearest Neighbour, the use of trees does not seem to have a major impact on the results when using Nearest Neighbour procedure. Therefore, as a general conclusion we can say that Frequency Distribution is the best performing method overall as it preserves marginal distributions, has a reasonable level of preservation of individual values, produces unbiased estimates for the total and has the lowest variability between all the methods.

Frequency Distribution and Nearest Neighbour methods produce unbiased estimates for the total number of records in a specific category. In contrast, the Highest Probability method does not lead to unbiased estimates as shown in the theoretical results.
In terms of variability, we can see that the values for the variances in the case of Frequency Distribution and Highest Probability methods are very similar. In contrast, Nearest Neighbour produces larger variances than the rest of the methods.
Comparisons between MSEs show that the lowest values are always found for Frequency Distribution followed by Nearest Neighbour and Highest Probability (due to the bias) respectively.

The results of the simulation confirm the theoretical result that the estimator of the variance proposed for the Frequency Distribution case is an unbiased estimator. However, even when in theory the estimator of the variance for the Nearest Neighbour seems to be unbiased, some differences between the real value and the estimator were found. However, these differences are probably big when estimating the variance but they are not very important in terms of the values of the point estimates as they are very small with respect to these values.

It has been shown in the results that the coverage, is over 94% all the time in the univariate case for both Frequency Distribution and Nearest Neighbour.

Comparisons between the proposed method and a Sequential Hot Deck method show that in terms of the point estimates any of the Frequency Distribution, Nearest Neighbour and Hot Deck produces unbiased estimators. In terms of variability, the sequential Hot Deck method produces larger variances than any of the imputation procedures investigated in this research. However, if a comparison between the mean square errors is made, we can see that sequential Hot Deck performs better than the Highest Probability procedure, producing smaller MSE.

Thus, Frequency Distribution is still the best performing imputation methods in this research, followed by Nearest Neighbour, Sequential Hot Deck and Highest Probability respectively.

CHAPTER 5

MULTIVARIATE CASE THEORETICAL FRAMEWORK

5.1. INTRODUCTION

This chapter extends *Chapter 3* by considering the case where more than one variable is subject to nonresponse.

Here, the multivariate case is explained including modelling description, the use of classification trees, imputation methods used and estimation of population quantities. Additionally, biases for the proposed estimator are studied.

5.2. NOTATION

Using the notation employed in *Chapter 3*, let U be a finite population of N elements $U=\{U_i; i=1,2,...,N\}$. Let $\mathbf{Y}=(y_{ih})$ be a $(NxH)-matrix$ of variables, where y_{ih} represents the $h\,th$ variable for the $i\,th$ element and let $\mathbf{X}=(x_{ik})$ be a $(NxK)-matrix$ of auxiliary variables where x_{ik} represents the $k\,th$ variable for the $i\,th$ element.

As also defined in *Chapter 3*, we now have $\mathbf{R}=(r_{ih})$ as the $(NxH)-matrix$ of indicator variables identifying whether or not y_{ih} is missing. That is, $r_{ih}=\begin{cases}1 & if\ y_{ih}\ is\ observed\\ 0 & otherwise\end{cases}$.

In this case, $\mathbf{Y}$ can be represented as $\mathbf{Y}=(\vec{Y}_1,\vec{Y}_2,...,\vec{Y}_h,...,\vec{Y}_H)$, where $\vec{Y}_h=(y_{1h},y_{2h},...,y_{Nh})^t$ is the vector of N values y_{ih}; $\mathbf{X}$ can be represented as $\mathbf{X}=(\vec{X}_1,\vec{X}_2,...,\vec{X}_k,...,\vec{X}_K)$, where $\vec{X}_k=(x_{1k},x_{2k},...,x_{Nk})^t$ is the vector of N values x_{ik}; and $\mathbf{R}$ can be represented as $\mathbf{R}=(\vec{R}_1,\vec{R}_2,...,\vec{R}_h,...,\vec{R}_H)$, where $\vec{R}_h=(r_{1h},r_{2h},...,r_{Nh})^t$ is a vector of N values r_{ih}.

It is assumed that each vector $\vec{Y}_h$ may be subject to nonresponse but that each vector $\vec{X}_k$ is fully observed. It is also assumed that the population is fully enumerated (no sample is taken).

The data takes the form:

	X_1	X_2	...	X_K	Y_1	Y_2	...	Y_H	R_1	R_2	...	R_H
1	x_{11}	x_{12}	...	x_{1K}	y_{11}	0	...	y_{1H}	1	0	...	1
2	x_{21}	x_{22}	...	x_{2K}	0	y_{22}	...	0	0	1	...	0
3	x_{31}	x_{32}	...	x_{3K}	y_{31}	0	...	y_{3H}	1	0	...	1
.	.	.	...	.	.	.	...	.	.	.	...	.
.	.	.	...	.	0	y_{n2}	...	y_{nH}	0	1	...	1
.	.	.	...	.	0	.	...	.	0	.	...	.
.	.	.	...	.	.	.	...	0	.	.	...	0
N	x_{N1}	x_{N2}	...	x_{NK}	y_{N1}	0	...	y_{NH}	1	0	...	1

where the zeros represent the missing values in the population and x_{ik} *and* y_{ih} are specific values for a specific realisation of the model. It is important to point out that in this case, the number of missing values can be different for each vector of variables $\vec{Y}_h$.

Additionally, we also define J_h as the number of categories for the variable y_{ih}. That is, variable y_{i1} has categories $j_1=\{1,2,...,J_1\}$; variable y_{i2} has categories $j_2=\{1,2,...,J_2\}$ and so on. In general we can say that variable y_{ih} has categories $j_h=\{1,2,...,J_h\}$.

5.3. MODEL DESCRIPTION

Under the model-based approach assumption, we consider x_{ik} and y_{ih} random variables with joint distribution $f(\mathbf{X},\mathbf{Y}\,|\,\boldsymbol{\theta})$ indexed by the vector of parameters $\boldsymbol{\theta}$.

As in the univariate case, the response process can be seen as a random process; therefore, the response outcome $\mathbf{R}$ is also included as matrix of random variables with distribution $f(\mathbf{R}\,|\,\mathbf{X},\mathbf{Y},\boldsymbol{\varphi})$.

Given that $\mathbf{X}$ is fully observed and $\mathbf{Y}$ is subject to nonresponse, we can now write the joint distribution as $f(\mathbf{X},\mathbf{Y},\mathbf{R}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})$ indexed by the vectors of parameters $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$.

The joint distribution of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{R}$, $f(\mathbf{X},\mathbf{Y},\mathbf{R}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})$, can be decomposed as the product of the probability distribution of $\mathbf{X}$ and $\mathbf{Y}$ indexed by vector of parameters $\boldsymbol{\theta}$ and the conditional distribution of $\mathbf{R}$ given $\mathbf{X}$ and $\mathbf{Y}$ indexed by $\boldsymbol{\varphi}$. That is,

$$f(\mathbf{X},\mathbf{Y},\mathbf{R}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}) \quad = \quad f(\mathbf{X},\mathbf{Y}\,|\,\boldsymbol{\theta}) \quad f(\mathbf{R}\,|\,\mathbf{X},\mathbf{Y},\boldsymbol{\varphi}) \qquad (1)$$

Since $\mathbf{Y}$ is subject to nonresponse, we can write $\mathbf{Y}=(\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}})$, where $\mathbf{Y}_{\mathbf{obs}}$, represents the observed part of $\mathbf{Y}$ and $\mathbf{Y}_{\mathbf{mis}}$, represents the missing part of $\mathbf{Y}$.

Therefore, the distribution $f(\mathbf{X},\mathbf{Y},\mathbf{R}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})$ can be written as $f(\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}},\mathbf{R}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})$.
Furthermore, equation (1) can be written as

$$f(\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}},\mathbf{R}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})=f(\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}}\,|\,\boldsymbol{\theta})f(\mathbf{R}\,|\,\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}},\boldsymbol{\varphi}).$$

The distribution of the observed data can be obtained by integrating $\mathbf{Y}_{\mathbf{mis}}$ out of the joint distribution of $\mathbf{X}$, $\mathbf{Y}\,and\,\mathbf{R}$. That is, $f(\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{R})=\int f(\mathbf{X},\mathbf{Y},\mathbf{R})\,d\mathbf{Y}_{\mathbf{mis}}$. More specifically, $f(\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{R}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})=\int f(\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}}\,|\,\boldsymbol{\theta})\,f(\mathbf{R}\,|\,\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}},\boldsymbol{\varphi})\,d\mathbf{Y}_{\mathbf{mis}}$.

Assumptions about the model are normally made in order to obtain valid estimation. One of the most common assumptions is that the missing values are "*missing at random*", MAR (Little and Rubin, 1987).

As in *Chapter 3*, the data is said to be missing at random if the response indicator $\mathbf{R}$ does not depend on the missing values of $\mathbf{Y}$, $\mathbf{Y}_{\mathbf{mis}}$. That is, MAR holds if $f(\mathbf{R}\,|\,\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\mathbf{Y}_{\mathbf{mis}},\boldsymbol{\varphi})=f(\mathbf{R}\,|\,\mathbf{X},\mathbf{Y}_{\mathbf{obs}},\boldsymbol{\varphi})$.

Then, assuming that MAR holds, and given that the actual observed data is $(\mathbf{X},\mathbf{Y}_{\text{obs}},\mathbf{R})$, we now have

$$f(\mathbf{X},\mathbf{Y}_{\text{obs}},\mathbf{R}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})=f(\mathbf{X},\mathbf{Y}_{\text{obs}}\,|\,\boldsymbol{\theta})f(\mathbf{R}\,|\,\mathbf{X},\mathbf{Y}_{\text{obs}},\boldsymbol{\varphi}).$$

Again, as in *Chapter 3*, the common maximum likelihood procedure used for fully observed data can be used for estimating the parameter $\boldsymbol{\theta}$ required when the data is incomplete (data with missing values). That is, $\boldsymbol{\theta}$ can be estimated using the maximum likelihood procedure over the observed data given that the missing data mechanism is ignorable, which means ignoring the second part of the right hand side of the last equation.

5.4. USING CLASSIFICATION TREES

As explained in *Chapter 3*, the use of classification trees for generating the imputation classes represents an important part of this research.
Also, as described in *Chapter 2* and 3, CART consists of grouping records depending on a set of values of the explanatory variables x_{ik}. The terminal nodes obtained from this classification are expected to be exclusive and exhaustive groups.

In practice, the explanatory variables can also be subject to nonresponse. However, we assume fully observed explanatory variables in this thesis.

As in *Section 3.3*, t represent the terminal nodes with $t \in \{1,2,...,t,...,T\}$, and T equal to the total number of terminal nodes for a specific tree. We also have a measurement vector $\bar{x}_i=(x_{i1},x_{i2},...,x_{iK})$ containing a number of measurements made on unit i. The collection of all possible measurement vectors defines the measurement space χ, with $\chi=\{\bar{x}_i\,;i=1,2,...,N\}$. We define χ_t as the set of measurement vectors belonging to a specific terminal node with $\chi=\chi_1 \cup \chi_2 \cup ... \cup \chi_T$.

Under the model-based assumption, we have the probability function of y_{ih} given the terminal node defined by χ_t as $f_h(y_{ih}=j\,|\,\bar{x}_i \in \chi_t)$. That is, the probability function of $\mathbf{Y}$ given a set of values of the explanatory variables identifying that terminal node. For simplicity, we write $f_h(y_{ih}=j\,|\,\bar{x}_i \in \chi_t)=f_h(y_{ih}=j\,|\,t)$.

Since all the variables used in this work are categorical, we denote $f_h(j_h \mid t) = P(y_{ih} = j \mid \vec{x}_i \in \chi_t)$ as the probability that y_{ih} takes the value j in the terminal node t, with $j = \{1,...,J_h\}$. Refer to *Section 3.3* in *Chapter 3* for an example.

The inclusion of the classification groups introduces a new factor to the distributions mentioned so far. Then, for a specific classification, we have $f_h(y_{ih} \mid \chi_t) = f_h(y_{ih} \mid t)$ as the probability function of y_{ih} given the terminal node t, and $f_h(y_{ih}, r_{ih} \mid t)$ as the joint distribution of y_{ih} and r_{ih} given the terminal node t. As in the model description (equation 1), the last equation can be decomposed as $f_h(y_{ih}, r_{ih} \mid t) = f_h(y_{ih} \mid t) f_h(r_{ih} \mid y_{ih}, t)$. Then, if MAR holds and assuming independence between units, $f_h(y_{ih} \mid r_{ih} = 0, t) = f_h(y_{ih} \mid r_{ih} = 1, t) = f_h(y_{ih} \mid t)$.

As in *Chapter 3*, since the imputation is done within terminal nodes we now assume MAR within terminal nodes. That is, $\mathrm{P}(y_{ih} = j_h) = f_h(j_h \mid t)$. See *Section 3.3* in *Chapter 3* for details.

5.5. COMPOSITE VARIABLE

In this work, the imputation process requires the generation of a classification tree as a first step. That classification tree is generally constructed for a single categorical variable. Therefore, as all the variables used in this work are basically categorical variables, one way to undertake joint imputation in the multivariate case is to create a variable that combines all the possible categories of the variables that are subject to nonresponse. This variable is called a "composite variable".

Then, a composite variable, denoted by y_i^c, is a variable that combines the values of all the possible variables subject to nonresponse, that is, y_i^c has categories $j = \{1, 2, ..., j^c, ...J^c\}$ with $J^c = J_1 x J_2 x ... x J_H$.

Example

To illustrate the construction of a composite variable let us suppose that we have three variables subject to missing information, y_{i1}, y_{i2} and y_{i3}. Suppose also that variable y_{i1} has two different categories, 1 and 2; and variables y_{i2} and y_{i3} have three different categories, 1, 2 and 3. Then, the composite variable will be a variable that contains 18 categories, as specified hereafter

Table 5.5.1
Example of categories of a composite variable

	Categories																	
y_{i1}	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
y_{i2}	1	1	1	2	2	2	1	1	1	2	2	2	1	1	1	2	2	2
y_{i3}	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Composite Var. (y_i^c)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**

It can be seen from *Table 5.5.1* that the composite variable is also a categorical variable containing 18 categories, which are the combinations of all possible categories of the variables subject to missing values. Then, a record with category 10 in the composite variable means a record with categories 2, 2 and 1 for the original variables y_{i1}, y_{i2} and y_{i3} respectively.

5.6. DIFFERENT APPROACHES TO IMPUTATION IN THE MULTIVARIATE CASE WHEN USING CLASSIFICATION TREES

Before embarking on a description of the different ways of using classification trees for imputing in a multivariate case, it is important to specify the different kind of situations that can be present in a multivariate case.

For simplicity, suppose from now on that we have just two vectors of random variables subject to nonresponse, $\vec{Y}_1$, $\vec{Y}_2$ where each y_{i1} can take categories $\{1,2,...,j_1,...,J_1\}$ and each y_{i2} categories $\{1,2,...,j_2,...,J_2\}$, and K fully observed vector of auxiliary variables $\vec{X}_k$. In this case, our composite variable y_i^c is a variable with categories $\{1,2,...,j^c,...,J^c\}$ where j^c corresponds to combination (j_1, j_2) and $J^c = J_1 x J_2$ is the number of categories for variable y^c.

There are three different combinations of missing information depending on which variable is missing at the time as shown in the next table

Table 5.6.1

Missing combinations for two variables

	Missing Variable	
Missing Combination	y_{i1}	y_{i2}
Combination 1	▨	
Combination 2		▨
Combination 3	▨	▨

It can be seen from the last table that combination 1 is the case in which only y_{i1} is missing, combination 2 the case where only y_{i2} is missing and combination 3 the case where both y_{i1} and y_{i2} are missing at the same time. This is a typical multivariate case, which involves missing values for more than one variable at the same time in the same database.

Let m be the number of records in the population for which both y_{i1} and y_{i2} are observed and N the size of that population. Let $a-m$ be the number of records with y_{i1} missing only, $b-a$ the number of records with y_{i2} missing only and $N-b$ the number of records with y_{i1} and y_{i2} missing at the same time. In this case, $N \geq b \geq a \geq m$. Then, the last table can be now written as follows

Table 5.6.2

Missing combinations and their respective total of cases within a tree

		Missing Variable	
Missing Combination	**Number of Cases**	y_{i1}	y_{i2}
Combination 1	$a-m$	▨	
Combination 2	$b-a$		▨
Combination 3	$N-b$	▨	▨

Additionally, as mentioned in *Section 5.4,* this research includes the use of classification trees as a first step within the imputation process, therefore, these combinations can be also found within terminal nodes. That is, for each terminal node we now have

Table 5.6.3

Missing combinations and their respective total of cases within terminal nodes

		Missing Variable	
Missing Combination	**Number of Cases**	y_{i1}	y_{i2}
Combination 1	$a_t - m_t$	▨	
Combination 2	$b_t - a_t$		▨
Combination 3	$N_t - b_t$	▨	▨

where $a_t - m_t$ is the number of records with y_{i1} missing only within a specific terminal node t; $b_t - a_t$ is the number of records with y_{i2} missing only in terminal node t; and $N_t - b_t$ is the number of records with y_{i1} and y_{i2} missing at the same time in terminal node t. Again, we have $N_t \geq b_t \geq a_t \geq m_t$.

There are different ways in which tree-based methods could be used for imputing values in a multivariate case. The imputation can be made using individual imputation, joint imputation or in a sequential imputation.

✓ In the first case, individual imputation, the imputation is done separately for each variable using separate trees. This would thus involve one tree for each variable missing in the analysis, without taking into account that they could be missing at the same time (see option 1 below). This approach would imply the use of different donors to impute different variables missing in the same record. Additionally, it could be very time consuming since many trees need to be generated and imputations are done separately.

✓ The second approach, a joint imputation approach, involves the imputation of all the missing values in a single record at the same time. Normally, these imputations are obtained from the same donor, that is, all the values missing in a specific recipient will be filled in using values coming from the same donor (see options 2, 3 and 4 below). In any of these options the imputation is done separately for each of the combinations described in *Table 5.6.3*. That means, either a classification tree is done for each combination in *Table 5.6.3* to be imputed (see options 2 and 3) or the imputations are carried out separately for those combinations but using the same tree (see option 4).
The main purpose of using joint imputation procedure is to preserve relationships between variables, that is, preserve joint distributions. This approach may also be faster since the donor is sought only once for each specific record.

✓ The last approach, a sequential approach, consists of imputing one variable at a time, but using that imputed information for the next step of the process. In this case, if two variables are missing at the same time for the same record, one of them is imputed first and then that imputed value is used as observed information either for growing a new tree for the other variable (see option 5) or for being used within the same tree for imputing the missing values of the other variable (see *Section 5.7*).
The sequential approach may also use different donors for filling in missing values in the same record and it can also be time consuming.
The difference between individual imputation and sequential imputation is that in the first case the imputations are done without taking into account extra information, while

in the second case the imputed values are used as observed information for the next imputation.

It is important to point out that in any case, either individual, joint or sequential imputation, the classification trees employed in the process can allow for missing covariates by the use of surrogates. However, we assume them to be fully observed in this thesis. An option in which missing information for the auxiliary variables can be used is presented in *Section 5.7.*

Additionally, these approaches take into account all the combinations of missing information present in the data, that is, all of the combinations 1, 2 or 3 mentioned in the two variables example presented in *Table 5.6.3* are assumed to occur at the same time.

Table 5.6.4 describes the possible options:

Table 5.6.4

Different options using tree models for imputing in the multivariate missing case

Option	Tree used	Action
1	individual tree for y_{i1} individual tree for y_{i2}	✓ Impute each variable y_{i1} and y_{i2} independently using the corresponding tree.
2	joint tree for y_{i1}, y_{i2} together. individual tree for y_{i1} individual tree for y_{i2}	✓ Impute missing values of y_{i1}, y_{i2} or (y_{i1}, y_{i2}) independently using the corresponding tree.
3	joint tree for y_{i1}, y_{i2} together.	✓ Impute (y_{i1}, y_{i2}) using the joint tree ✓ Impute y_{i1} and y_{i2} using an extension of the joint tree. ✓ Repeat the process for the other variable, say y_{i1} using now y_{i2} for expanding the tree.
4	joint tree for y_{i1}, y_{i2} together.	✓ Impute all y_{i1}, y_{i2} and (y_{i1}, y_{i2}) using the same tree
5	individual tree for y_{i1} individual tree for y_{i2} including y_{i1} as a complete variable	✓ Impute y_{i1} using the corresponding tree ✓ Impute y_{i2} using the correspondent tree, which includes the values of y_{i1} already imputed

A more detailed description of these options is now given.

Option 1

In this case imputations are carried out individually using one tree for each variable missing. For example, if two variables y_{i1} and y_{i2} are missing separately and together as explained before, all the records with y_{i1} missing will be imputed using the tree grown for that variable (y_{i1}) and all records with variable y_{i2} missing will be also imputed using the tree created for y_{i2}. The records for which the two variables are missing at the same time will be imputed in the same way, all y_{i1} using the tree for y_{i1} and all the y_{i2} missing using the tree for y_{i2}. Since, this approach uses different donors to impute different variables missing in the same record, joint distributions may not be preserved. Additionally, it is more time consuming than other options since the imputations are done separately using different trees, and records are imputed one at the time.

Option 2

This option implies the use of a different imputation tree for each missing combination in *Table 5.6.3*. That means, three different trees for imputation will be used. One tree for imputing y_{i1} alone, one tree for imputing y_{i2} alone and one tree for imputing the combination of two of them, (y_{i1}, y_{i2}). In the last case, the imputation is done throughout a composite variable formed by the all possible combinations between y_{i1} and y_{i2} as explained in *Section 5.5*.

This option has the advantage that the classification is especially created for the combination missing (either for a single variable or for a combination of many variables). This allows for the use of more accurate classification for each combination missing. However, this procedure can be computer intensive since the number of trees required increases with the number of missing variables (or combinations).

Option 3

This option comprises the use of one classification tree for imputing all the missing combinations but expanding the terminal nodes depending on the variable being missing. For example, in the case of two variables subject to missing information shown in *Table 5.6.3*, the process followed is,

1. grow a tree for the combination missing, joint tree for (y_{i1}, y_{i2})
2. use that tree impute the combination (y_{i1}, y_{i2}) missing
3. using the observed information for y_{i1} expand the tree further to a new set of terminal nodes for the other variable missing y_{i2} and then impute that variable y_{i2} using the new set of terminal nodes obtained
4. as above, using the original tree generated for the combination (y_{i1}, y_{i2}) and using the observed information for y_{i2} expand that original tree further to a new set of terminal nodes for the other variable missing y_{i1} and then impute that variable y_{i1} using the new set of terminal nodes

This option has the advantage that only one tree is used, even when this is expanded. However, the extension made implies a new procedure that can be more time consuming than others since each terminal node of the original tree (tree for (y_{i1}, y_{i2})) is now used as a new database for expanding it.

Option 4

Sometimes it may be a very difficult task to grow trees for each combination of missing variables because of complexity or the time taken. Additionally, combinations (composite variables) involving many variables may require the use of too many categories making the analysis also more complex and more time consuming as well. An option to solve this problem can be the use of one tree to impute all the possible combinations missing in a data set. The selection of the tree to be used will depend on which combinations are missing. One possibility is to use the combination with the largest percentage of missing information.

This option has the advantage of using joint imputation since all the missing values in a recipient are filled in using values from the same donor. However, the classification used for imputation may be right for some combinations but not for others. It will depend on how related the variables are.

Example

To illustrate this point, suppose that we have four different variables with three different combination of missing values, say: 1) y_{i1} and y_{i2}, 2) y_{i1}, y_{i2}, and y_{i3}, and 3) y_{i1}, y_{i2}, and y_{i4}. Suppose also that the percentage of missing information is as follows: 10% for the first combination, 3% for the second combination and 2% for the last combination. In this case, it could be complicated and time consuming to grow trees for all of the three combinations. Instead, it may be easier to grow a tree for the combination number 1 (y_{i1} and y_{i2}) as this combination has the largest percentage of missing information and is also present on the rest of the missing combinations, and then use this tree for imputing. The imputation can be made either jointly for y_{i1} and y_{i2} in all the records in combinations 1, 2 and 3 and then individually for the rest of variables not imputed yet y_{i3} and y_{i4}; or for all the variables missing in all the combinations, say y_{i1} and y_{i2} for the first combination; y_{i1}, y_{i2}, and y_{i3} for the second combination; and y_{i1}, y_{i2}, and y_{i4} for the last one; but using the same classification tree generated at the beginning.

Additionally, any value missing individually (e.g. y_{i3} alone or y_{i4}) can be imputed using the same tree.

In any case, the task of growing trees is reduced from three trees to one tree. This option can be useful for combinations with very small percentage of missing information that involve very large number of variables or categories.

Option 5

One way of using a sequential procedure in the case of two missing variables is to grow a tree for a single missing variable (say y_{i1}) and impute that variable. Once that variable is imputed, its value may be used as a complete covariate for growing the tree for the other single missing variable, y_{i2} and then impute that variable. In this case, the process may become slow since many trees have to be created (as many as missing variables are). Another disadvantage of this method is again the fact that the imputations for the same record come from different donors, which can make the preservation of joint distributions a more difficult task.

Additionally, there should be a pre-established order for the variables to be imputed, that is, which variable is imputed first, which is imputed second and so on.

5.7. SPECIAL CASE, USING AUXILIARY VARIABLES WITH MISSING VALUES

Up to this point, we have been assuming that all auxiliary variables are fully observed, however, sometimes that is not necessarily true. The case where missing information is present not only for the target variable but also for the auxiliary information is very common. In this sense, the following procedure represents an alternative when using classification trees for imputing in this kind of cases.

A way to use CART for imputation in the presence of missing covariates is, once a missing value in a covariate is found, use a surrogate to classify it, impute it and keep going down the tree. That is, once a tree is created using fully observed records, if the variable for classification (variable x_{ik}) is missing in a specific recipient, that record is classified by using a surrogate variable and imputed immediately using information from the node in which this is. Then, after imputing that value, the classification process continues, following the same procedure. Once the tree has reached the terminal nodes, the imputation for the target variable is carried out as in any of the cases mentioned in the last section.
The definition and selection of surrogates is explained in *Section 2.3.5* in *Chapter 2*.

Example

To illustrate this approach, let us consider the following example. Suppose there is one target variable y_i^c of interest (for which estimation is required) which is a composite variable created by the combination of two different variables y_{i1} and y_{i2} missing at the same time and four different independent explanatory variables x_{ik} with the following categories

Table 5.7.1

Categories for an example of a set of variables

Variable	Categories
y_i^c	1,2,3,4,5,6
x_{i1}	1,2
x_{i2}	1,2,3
x_{i3}	1,2,3,4
x_{i4}	0,1

Additionally, suppose that the records with variable y_i^c missing are as follows

Table 5.7.2

Example of a database with missing values

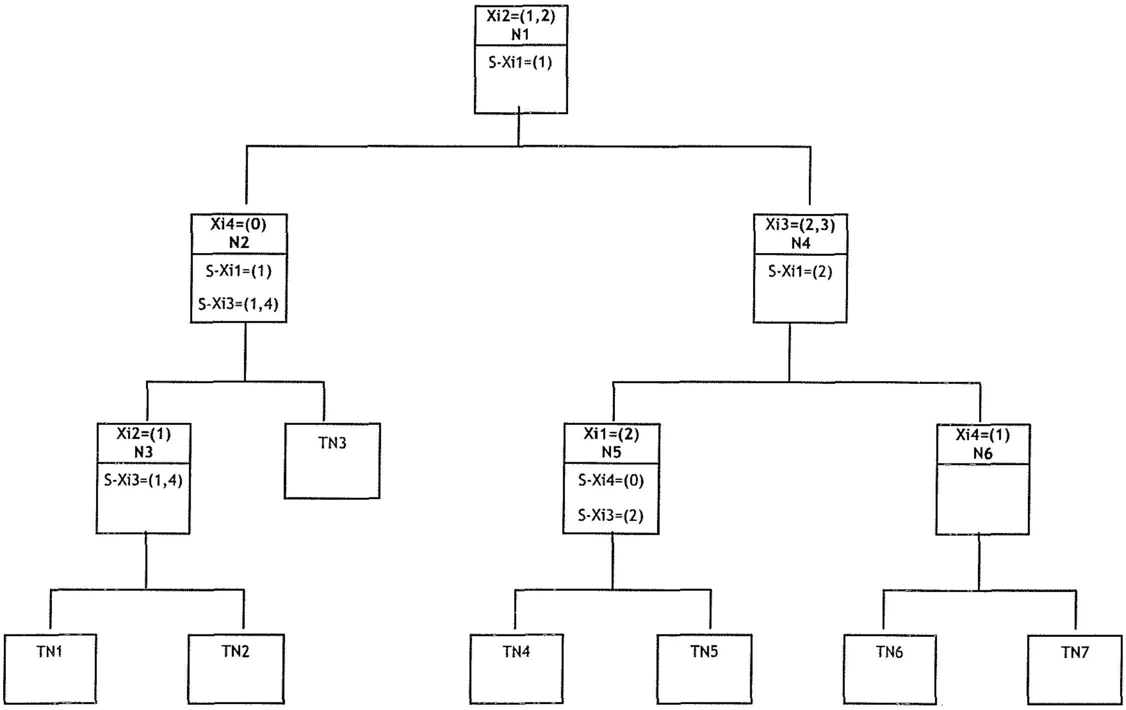
Records	y_i^c	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	missing	1	missing	4	0
2	missing	2	1	3	missing
3	missing	missing	3	2	1

It can be noticed from the last table that missing values are present not only at the target variable y_i^c but also in some of the independent variables used in the analysis for generating the tree.

Suppose also that the corresponding classification tree generated using only observed x_{ik} is as follows

Figure 5.7.1

Example of a Classification Tree for its use in a Sequential Imputation



It is important to point out that the tree can be generated using also incomplete covariates, however, only observed information was used in this example in order to simplify the process.

In this case, the classification tree is grown for the variable y_i^c based only on fully observed variables x_{ik}. Once the tree is generated, the first node, N1, is defined by values 1 and 2 of variable x_{i2}, that is, all the records with values 1 or 2 for variable x_{i2} go to the left node, N2, and records with any other values for variable x_{i2} (x_{i2} =3 in this case) go to the right node, N4. If variable x_{i2} is missing, which is the case of record one of this example, that record will be classified using the surrogate available in this case, which is $s-x_{i1}=(1)$ (see

Section 2.3.5 in *Chapter 2* for definition and selection of surrogates). That means, if x_{i2} is missing, the record will go to the left node, N2, if $x_{i1} = (1)$ and to the right node, N4 otherwise ($x_{i1} = 2$). Once the record with missing values for x_{i2} is classified, the value of x_{i2} will be imputed using a value from a donor chosen from the node where that record ends up (N2 in this case, as its value for x_{i2} is missing and $x_{i1} = (1)$). That is, the imputation will be one of the values of x_{i2} present in node 2 depending on the modal category of that node. Following the same line, for the next classification step on the same record is to look at the value of x_{i4}. In this example $x_{i4} = 0$, then record one will go to N3 and then, depending on the imputed value for x_{i2}, record one will end up either in TN1 or TN2.

Once the record is successfully classified, the donor for imputing the y_i^c variable missing will be chosen from the correspondent terminal node where the recipient ends up, which are TN1 or TN2 in this case. The same procedure is followed for each of the recipients.

The disadvantage of this approach is again the fact of using different donors to impute different missing items in the same recipient, risking the maintenance of the joint distribution, which is an important aspect to be considered when analysing census (or any) data.

5.8. OPTION CONSIDERED IN THIS RESEARCH

As mentioned in *Section 1.8* in *Chapter 1*, one of the aims of this thesis is to develop a joint imputation procedure, which maintains the joint distribution as much as possible. In order to achieve that, a good imputation procedure to implement could be a joint imputation process with the imputated values coming from the same donor, guaranteeing the preservation of the joint distribution since the relationship between variables is not distorted.
Additionally, since most of the variables used in this thesis are categorical variables, the use of composite variables seems to be reasonable and easy to implement.

The approach undertaken in this work is then the use of classification trees in conjunction with joint imputation using composite variables since it seems to be reasonable good in efficiency, that is, in time consumed and precision of the results. Therefore, the option implemented in this work is Option 4 where a joint tree is used for imputing the variables missing together.

Additionally, in cases where the pattern of missing information is very large and complex, the approach in this thesis will tend to be the construction of classification trees for the most important combinations in order to use these trees for imputing combinations with low percentages of missing information and large number of variables, as described in Option 4 as well. In practice, this represents a feasible and good option since the classification can be carried out in the presence of missing covariates by the use of surrogates and only one tree is required. However, due to time constraints, this option is not further investigated in this research.

Option 4 can be divided into two different alternatives: when many variables are missing at the same time but not individually (each of them separately) and when the variables are missing at the same time and also individually.

In a two missing variables case, the first alternative is carried out using a composite variable for the combination of the missing variables (combination 3 in *Table 5.6.3*) as is shown in the example presented in *Section 5.5*. After this variable is created, a classification tree is grown as if it was a single variable with number of categories equal to the number of possible combinations of the single variables involved in the process. Since the imputation is carried out for individual categories of a variable (composite variable), the properties of the estimators proposed when using single imputation approach together with classification trees shown in the univariate case are still valid when using a composite variable (see *Section 5.11*).

The second alternative uses the joint tree for imputing not only the combination of the missing variables, but also the individual variables missing alone (combinations 1,2 and 3 in *Table 5.6.3*). In this case, the process is carried out using the same tree for every imputation, as in the last case, but as many times as there are missing combinations. For example, in the case presented in *Table 5.6.3,* the joint tree for the combination (y_{i1}, y_{i2}) will be used for imputing the records with that combination missing at the same time as well as for imputing the records with the individual variables missing alone.

5.9 IMPUTATION METHODS

As in the univariate case, each imputation method is applied at each terminal node. The same three different imputation methods used in the univariate case are considered here.

In order to explain the imputation methods used, let us assume that the variable of interest is a composite variable. That is, in a two variables case, we can have combinations 1,2 and 3 presented in *Table 5.6.3* as explained in *Section 5.6.*

In this case, we have that category j^c of variable y_i^c correspond to combination (j_1, j_2) for variables y_{i1} and y_{i2} respectively. Additionally, we define $\hat{y}_i^c$ to be the imputed value of variables y_i^c obtained by any of the imputation methods. That is, $\hat{y}_i^c$ can be $(\hat{j}_1, j_2)$ for combination 1 where only variable y_{i1} is missing, $(j_1, \hat{j}_2)$ for combination 2 where only variable y_{i2} is missing and $(\hat{j}_1, \hat{j}_2)$ for combination 3 where both variables y_{i1} and y_{i2} are missing at the same time, as shown in *Table 5.6.3*.

The explanatory variables x_{ik} are still considered fully observed in this section.

1. Probability Distribution Method

As in any case with missing values, we want to impute y_i^c when this is missing from $f(y_i^c \mid r_i^c = 0, t, \theta)$.

Assuming that MAR holds and assuming independence between the units, we can write $f\left(y_i^c \mid r_i^c = 0, \vec{x}_i\right) = f\left(y_i^c \mid r_i^c = 1, \vec{x}_i\right)$.

For a tree model it is supposed that $f(y_i^c \mid r_i^c = 0, t, \boldsymbol{\theta}) = f(y_i^c \mid r^c{}_i = 1, t, \boldsymbol{\theta})$, where $f(y_i^c \mid r_i^c = 1, t, \boldsymbol{\theta})$ is the probability distribution of the observed values given the terminal node t, $f(y_i^c \mid r^c{}_i = 0, t, \boldsymbol{\theta})$ is the probability distribution of the missing values given the terminal node t. Then, the probability distribution method works as follows: given a specific tree, and for each terminal node we use the estimated distribution of the observed variables $f(y_i^c \mid r_i^c = 1, t, \boldsymbol{\theta})$, i.e. $P(y_i^c \mid r_i^c = 1, t)$ for imputing the missing values.

Since variable y_i^c is categorical, we write $\mathrm{P}(y_i^c = j^c \mid r_i^c = 1, t, \boldsymbol{\theta}) = p_{j^c t}$, where j^c represents the categories of the variable y_i^c with $j^c \in \left\{1, 2, \ldots, J^c\right\}$.

The probability $p_{j^c t}$ can be estimated in three different ways depending on the missing combination. That is,

1. If the missing combination is combination 1 in *Table 5.6.3,* $p_{j^c{}_t}$ is estimated by the observed proportion of cases with category j^c of variable y_i^c given $y_{i2} = j_2$, $\hat{p}_{j^c/j_2t} = \sum_{i=1}^{m_t} \frac{I(y_i^c = j^c)}{m_{j_2t}}$, with m_{j_2t} as the number of observed records with $y_{i2} = j_2$.
2. If the missing combination is combination 2 in *Table 5.6.3,* $p_{j^c{}_t}$ is estimated by the observed proportion of cases with category j^c of variable y_i^c given $y_{i1} = j_1$, $\hat{p}_{j^c/j_1t} = \sum_{i=1}^{m_t} \frac{I(y^c = j^c)}{m_{j_1t}}$, with m_{j_1t} as the number of observed records with $y_{i1} = j_1$.
3. If the missing combination is combination 3 in *Table 5.6.3,* $p_{j^c{}_t}$ is estimated by the observed proportion of cases with category j^c of variable y_i^c, $\hat{p}_{j^c{}_t} = \sum_{i=1}^{m_t} \frac{I(y^c = j^c)}{m_t}$, with m_t as the number of observed records.

In other words, the probability distribution of y_i^c for the missing data is assumed to be equal to the probability distribution of y_i^c for the observed data, which is estimated by any of the options mentioned before depending on the combination missing.

2. Highest Probability Method (or Modal Imputation)

Under the same assumptions made for the probability distribution method, that is, MAR holds and independence between units, and given a specific tree, this method imputes the value that is "most likely" in that specific terminal node (i.e. has the highest probability) to all of the records with missing values. Thus, the value to be imputed will be j^{c*} and can be divided in three cases

1. If the missing combination is combination 1 in *Table 5.6.3,* the value to be imputed will be j^{c*} satisfying $\hat{p}_{j^{c*}{}_t} \geq \hat{p}_{j^c/j_2t}$, for all categories j^c of the response variable.
2. If the missing combination is combination 2 in *Table 5.6.3,* the value to be imputed will be j^{c*} satisfying $\hat{p}_{j^{c*}{}_t} \geq \hat{p}_{j^c/j_1t}$, for all categories j^c of the response variable.
3. Finally, if the missing combination is combination 3 in *Table 5.6.3,* the value to be imputed will be j^{c*} satisfying $\hat{p}_{j^{c*}{}_t} \geq \hat{p}_{j^c{}_t}$, for all categories j^c of the response variable.

Then, in this case, the imputation takes the value $\hat{y}_i^c = j^{c*}$

It could be more than one j^{c*} value satisfying this condition. In this case, the method selects one of the categories randomly with equal probabilities.

3. Nearest Neighbour Method

Given a specific tree and for each terminal node individually, distances between the recipient and each possible donor are calculated and the "nearest" donor defines the imputed value for that particular recipient. The nearest donor is determined by the set of independent variables. That is, the distance between the two records (recipient and possible donor) is calculated by adding one to the distance function every time different values are found between them for the independent variables.

Then, given a recipient i' with values $x_{i'k}$, $k = 1,2,...,K$ for the variables x_{ik}, a donor i with value y_i^c for the variable y_i^c and values x_{ik}, $k = 1,2,...,K$ for the variables x_{ik} is that record which satisfies $\min_i [d_{i'i}]$ with $i' \in nr_t^c$ and $i \in r_t^c$, and $d_{i'i} = \sum_{k=1}^{K} I(x_{i'k} \neq x_{ik})$

Then, the missing value $y_{i'}^c$ will be imputed with the observed value y_i^c from the donor i, $\hat{y}_{i'}^c = y_i^c$. As explained before, $\hat{y}_{i'}^c$ can be $(\hat{j}_1, j_2)$ for combination 1 as only variable y_{i1} is missing, $(j_1, \hat{j}_2)$ for combination 2 as only variable y_{i2} is missing and $(\hat{j}_1, \hat{j}_2)$ for combination 3 as both variables y_{i1} and y_{i2} are missing at the same time.

In this case we define A_i as the number of times unit i is used as donor, therefore, $A_i = \sum_{i \in nr_t^c} I(d_{i'i} \leq d_{i'l} \; for \; all \; l \in r_t^c)$.

It is important to point out that in this case a record can be used more than once as a donor. This means that if a specific record has the least distance to two different recipients, this record could be used as a donor to fill in the missing values for both of the recipients. Moreover, when a recipient has the same distance to two different donors, one of the donors is randomly selected with equal probabilities.

5.10 ESTIMATION OF POPULATION QUANTITIES

As in the univariate case, the estimation will be concentrated in population quantities rather than superpopulation parameters.

As mentioned in *Section 5.6* we will concentrate in the two missing variables case, that is, only two target variables are subject to nonresponse and the rest of the variables, which are auxiliary variables, are assumed to be fully observed.

In order to simplify the estimation procedure, we divide combinations in *Table 5.6.3* into two different cases. We call **Case 1** the case where the two variables are missing at the same time and they are not missing on their own, combination 3 in *Table 5.6.3*. This is a simple case to treat since it can be seen as a univariate case where a composite variable is used. In the same context, we call **Case 2** the case where any of the three combinations mentioned before can be present in the data, that is, the two variables can be missing individually and together. This case requires a more complex formulation since we need to include all the possible missing combinations in the estimation procedure.

Since we have two variables involved in the missing process, we may be interested in two different estimators. First, we may want to estimate the total of cases with category j_1 of the variable y_{i1} or equivalently, total of cases with category j_2 of the variable y_{i2}. Second, we can consider the estimator of the total of cases with category j_1 of variable y_{i1} and category j_2 of variable y_{i2} together, that is, the total of cases with category j^c of the variable y_i^c. However, we will concentrate in the second option, where the parameter of interest is the total of cases with category j_1 of variable y_{i1} and category j_2 of the variable y_{i2} together.

5.10.1. CASE 1 - Variables missing at the same time only

Using the notation described in *Section 5.2,* we have a finite population $U=\{U_i; i=1,2,...,N\}$ of N elements and a variable of interest y_i^c, $i=1,2,...,N$. The aim is to estimate a population quantity, for example, the total $Y^c=\sum y_i^c$.

It can be noticed that in this case the number of missing cases for some combinations in *Table 5.6.3* are zero, that is, $a_t - m_t = 0$, $b_t - a_t = 0$ and $N_t - b_t \neq 0$, in which case we have $a_t = b_t = m_t$.

Given that the quantity of interest is the total of cases i with category j^c for the variable y_i^c, which in our case is a categorical variable taking values $j^c = \{1,...,J^c\}$, this total can be written as $g_{j^c} = \sum_U I(y_i^c = j^c)$.

Since not all the data is observed, this population quantity can be estimated as follows $\hat{g}_{j^c} = \sum_{i=1}^{m_t} I(y_i^c = j^c) + \sum_{i=m_t+1}^{N_t} I(\hat{y}_i^c = j^c)$, where $\hat{y}_i^c$ is the imputed value of the variable y_i^c for the unit i if missing.

The first part of the expression, $\sum_{i=1}^{m_t} I(y_i^c = j^c)$, can be obtained from the observed data, while the second part of the expression, $\sum_{i=m_t+1}^{N_t} I(\hat{y}_i^c = j^c)$ is determined by the imputation method used.

5.10.2. CASE 2 - Variables missing individually and together

As before, the aim is to estimate the same population quantity, total of cases i with category j^c for the variable y_i^c. In this case, all the records imputed either individually (only one variable missing at the time) or jointly (two variables missing at the same time) must be included in the total calculation.

It can be noticed that in this case the number of missing cases for all of the combinations in *Table 5.6.3* are different from zero, that is, $a_t - m_t \neq 0$, $b_t - a_t \neq 0$ and $N_t - b_t \neq 0$, in which case we have $a_t \neq m_t$ and $a_t \neq b_t$. That is, the total of records imputed as category j^c will include records with category j_1 imputed and category j_2 observed, the records with category j_1 observed and category j_2 imputed, and records with categories j_1 and j_2 imputed.

The number of cases in category j^c of y_i^c can be represented as $g_{j^c} = \sum_U I((y_{i1}, y_{i2}) = (j_1, j_2))$, where $(j_1, j_2) = j^c$ or in a similar way, $g_{j^c} = \sum_U I(\vec{y}_i = (j_1, j_2))$ where $\vec{y}_i = (y_{i1}, y_{i2})$.

As before, we distinguish between the three different combinations mentioned in *Table 5.6.3*. That is, given that $\vec{y}_i = (y_{i1}, y_{i2}) = (j_1, j_2)$ we then have three different kinds of imputed records. The first is when only y_{i1} is missing, in which case $\hat{\vec{y}}_i = (\hat{y}_{i1}, y_{i2}) = (j_1, j_2)$, the second case is when only y_{i2} is missing, in which case $\hat{\vec{y}}_i = (y_{i1}, \hat{y}_{i2}) = (j_1, j_2)$ and the third case is when both variables y_{i1} and y_{i2} are missing at the same time, in which case $\hat{\vec{y}}_i = (\hat{y}_{i1}, \hat{y}_{i2}) = (j_1, j_2)$.

Since not all the data is observed, this population quantity can be estimated as follows

$$g_{j^c} = \sum_{i=1}^{m} I(\vec{y}_i = (j_1, j_2)) + \sum_{i=m+1}^{a} I\left(\hat{\vec{y}}_i = (j_1, j_2)\right) + \sum_{i=a+1}^{b} I\left(\hat{\vec{y}}_i = (j_1, j_2)\right) + \sum_{i=b+1}^{N} I\left(\hat{\vec{y}}_i = (j_1, j_2)\right)$$

or in a similar way, $\hat{g}_{j^c} = \sum_{i=1}^{m_t} I(y_i^c = j^c) + \sum_{i=m_t+1}^{N_t} I(\hat{y}_i^c = j^c)$, where $\hat{y}_i^c$ is the imputed value of the variable y_i^c for the unit i if missing.

Again, the first part of the expression, $\sum_{i=1}^{m_t} I(y_i^c = j^c)$, is based on the observed data, while the second part of the expression, $\sum_{i=m_t+1}^{N_t} I(\hat{y}_i^c = j^c)$ is determined by the imputation method used.

5.11 PROPERTIES OF THE ESTIMATORS

Let us examine the bias properties of the total estimator presented in the last section. To avoid complexity, we shall not attempt to consider variance properties as in *Chapter 3*.

We are assuming model-based approach, where $P^{\xi}(y_i^c = j^c \mid \chi_t)$, $P^{\xi}(y_{i1} = j_1 \mid \chi_t)$ and $P^{\xi}(y_{i2} = j_2 \mid \chi_t)$ represents the probability that y_i^c takes the value j^c in a specific terminal node t given by the model ξ, probability that y_{i1} takes the value j_1 in a specific terminal node t given by the model ξ and probability that y_{i2} takes the value j_2 in a specific terminal node t given by the model ξ respectively.

We are also assuming that variable y_i^c, y_{i1} and y_{i2} are missing at random within terminal nodes, that means, $\mathrm{P}^{\xi}(y_i^c = j^c \mid t, r_i^c = 1) = \mathrm{P}^{\xi}(y_i^c = j^c \mid t, r_i^c = 0)$, $\mathrm{P}^{\xi}(y_{i1} = j_1 \mid t, r_{i1} = 1) = \mathrm{P}^{\xi}(y_{i1} = j_1 \mid t, r_{i1} = 0)$ and $\mathrm{P}^{\xi}(y_{i2} = j_2 \mid t, r_{i2} = 1) = \mathrm{P}^{\xi}(y_{i2} = j_2 \mid t, r_{i2} = 0)$ and holding r_i^c, r_{i1} and r_{i2} and x_{ik} as fixed.

5.11.1. Probability Distribution Imputation Case

Case 1

In this case, y_{i1} and y_{i2} are missing at the same time and not individually, that is, combination 3 in *Table 5.6.3.*

Our interest is to obtain an estimator for the total g_{j^c} of cases with category j_1 of y_{i1} and category j_2 of y_{i2}. In this case, we can use the estimator $\hat{g}_{j^c}$ defined in *Section 5.10.1.*

The bias of the estimator is given by

$$E_{\xi I}\left[\hat{g}_{j^c} - g_{j^c}\right] = E_{\xi I}\left[\sum_t \left[\sum_{i=1}^{m_t} I(y_i^c = j^c) + \sum_{i=m_t+1}^{N_t} I(\hat{y}_i^c = j^c)\right] - \sum_t \sum_{i=1}^{N_t} I(y_i^c = j^c)\right]$$

It can be noticed that this case can be treated as univariate case where the missing variable is a single variable y_i^c with categories $\{1, 2, \ldots, j^c, \ldots, J^c\}$. Therefore, the results are the same as in the univariate case (see *Section 3.6.1* in *Chapter 3*).

Case 2

In this case, variables y_{i1} and y_{i2} can be missing at the same time and also individually. That is, combination 1, combination 2 and combination 3 together.

As in case 1, we want to estimate the total of cases with categories j_1 *and* j_2 of variables y_{i1} and y_{i2} respectively, that is, the number g_{j^c} of cases with category j^c of y_i^c. We use the estimator defined in *Section 5.10.2*. The bias of this estimator is given by the expected value of the difference between the estimator and the parameter.

Let us define category (j_{10}, j_{20}) as the specific category to be estimated. So, we would like to estimate the number of records for which $\vec{y}_i = (j_{10}, j_{20})$

$$E_{\xi I}\left[\hat{g}_{j^c} - g_{j^c}\right] = E_{\xi I}\left[\sum_t\left[\sum_{i=1}^{m_t} I(y_i^c = j^c) + \sum_{i=m_t+1}^{N_t} I(\hat{y}_i^c = j^c)\right] - \sum_t\sum_{i=1}^{N_t} I(y_i^c = j^c)\right]$$

$$= E_{\xi I}\left[\sum_t\left[\sum_{i=1}^{m_t} I(\vec{y}_i = (y_{i1}, y_{i2})) + \sum_{i=m_t+1}^{a_t} I(\hat{\vec{y}}_i = (\hat{y}_{i1}, y_{i2})) + \sum_{i=a_t+1}^{b_t} I(\hat{\vec{y}}_i = (y_{i1}, \hat{y}_{i2}))\right]\right]$$

$$+ E_{\xi I}\left[\sum_t\left[\sum_{i=b_t+1}^{N_t} I(\hat{\vec{y}}_i = (\hat{y}_{i1}, \hat{y}_{i2})) - \sum_{i=1}^{N_t} I(\vec{y}_i = (y_{i1}, y_{i2}))\right]\right]$$

$$= E_{\xi I}\left[\sum_t\left[\sum_{i=m_t+1}^{a_t} I(\hat{\vec{y}}_i = (\hat{y}_{i1}, y_{i2})) + \sum_{i=a_t+1}^{b_t} I(\hat{\vec{y}}_i = (y_{i1}, \hat{y}_{i2})) + \sum_{i=b_t+1}^{N_t} I(\hat{\vec{y}}_i = (\hat{y}_{i1}, \hat{y}_{i2}))\right]\right]$$

$$- E_{\xi I}\left[\sum_t\left[\sum_{i=m_t+1}^{N_t} I(\vec{y}_i = (y_{i1}, y_{i2}))\right]\right]$$

$$= E_{\xi I}\left[\sum_t\left[\sum_{i=m_t+1}^{a_t} I(\hat{j}_{i1} = j_{10}) I(j_{i2} = j_{20}) + \sum_{i=a_t+1}^{b_t} I(j_{i1} = j_{10}) I(\hat{j}_{i2} = j_{20})\right]\right]$$

$$+ E_{\xi I}\left[\sum_t\left[\sum_{i=b_t+1}^{N_t} I(\hat{j}_{i1} = j_{10}) I(\hat{j}_{i2} = j_{20}) - \sum_{i=m_t+1}^{N_t} I(j_{i1} = j_{10}) I(j_{i2} = j_{20})\right]\right]$$

$$= E_{\xi}\left[\sum_t\left[\sum_{i=m_t+1}^{a_t} E_I\left[I(\hat{j}_{i1} = j_{10})\right] I(j_{i2} = j_{20}) + \sum_{i=a_t+1}^{b_t} I(j_{i1} = j_{10}) E_I\left[I(\hat{j}_{i2} = j_{20})\right]\right]\right]$$

$$+ E_{\xi}\left[\sum_t\left[\sum_{i=b_t+1}^{N_t} E_I\left[I(\hat{j}_{i1} = j_{10}) I(\hat{j}_{i2} = j_{20})\right] - \sum_{i=m_t+1}^{N_t} I(j_{i1} = j_{10}) I(j_{i2} = j_{20})\right]\right]$$

assuming independence between units

$$= E_{\xi}\left[\sum_t\left[\sum_{i=m_t+1}^{a_t} \hat{p}_{(j_{10}, j_{20})/j_{20}t} I(j_{i2} = j_{20}) + \sum_{i=a_t+1}^{b_t} \hat{p}_{(j_{10}, j_{20})/j_{10}t} I(j_{i1} = j_{10}) + (N_t - b_t)\hat{p}_{j_{10}t}\hat{p}_{j_{20}t}\right]\right]$$

$$- E_{\xi}\left[\sum_t\sum_{i=m_t+1}^{N_t} I(j_{i1} = j_{10}) I(j_{i2} = j_{20})\right]$$

$$= E_{\xi}\left[\sum_{t}\left[\hat{p}_{(j_{10},j_{20})/j_{20}t}\sum_{i=m_t+1}^{a_t} I(j_{i2}=j_{20})+\hat{p}_{(j_{10},j_{20})/j_{10}t}\sum_{i=a_t+1}^{b_t} I(j_{i1}=j_{10})+\sum_{t}\left[(N_t-b_t)\hat{p}_{j_{10}t}\hat{p}_{j_{20}t}\right]\right]\right]$$

$$-E_{\xi}\left[\sum_{t}\sum_{i=m_t+1}^{N_t} I(j_{i1}=j_{10})I(j_{i2}=j_{20})\right]$$

$$= \left[\sum_{t}\left[E_{\xi}\left[\hat{p}_{(j_{10},j_{20})/j_{20}t}\right]E_{\xi}\left[\sum_{i=m_t+1}^{a_t} I(j_{i2}=j_{20})\right]+E_{\xi}\left[\hat{p}_{(j_{10},j_{20})/j_{10}t}\right]E_{\xi}\left[\sum_{i=a_t+1}^{b_t} I(j_{i1}=j_{10})\right]\right]\right]$$

$$+\sum_{t}\left[\left[(N_t-b_t)E_{\xi}\left[\hat{p}_{j_{10}t}\right]E_{\xi}\left[\hat{p}_{j_{20}t}\right]\right]-\sum_{i=m_t+1}^{N_t} E_{\xi}\left[I(j_{i1}=j_{10})I(j_{i2}=j_{20})\right]\right]$$

$$= \sum_{t}\left[P_{j_{10}/j_{20}t}{}^{\xi}(a_t-m_t)P_{j_{20}}{}^{\xi}+P_{j_{20}/j_{10}t}{}^{\xi}(b_t-a_t)P_{j_{10}}{}^{\xi}+(N_t-b_t)P_{j_{10}t}{}^{\xi}P_{j_{20}t}{}^{\xi}-(N_t-m_t)P_{j_{10}t}{}^{\xi}P_{j_{20}t}{}^{\xi}\right]$$

$$= \sum_{t}\left[P_{j_{10}}{}^{\xi}(a_t-m_t)P_{j_{20}}{}^{\xi}+P_{j_{20}}{}^{\xi}(b_t-a_t)P_{j_{10}}{}^{\xi}+(N_t-b_t)P_{j_{10}t}{}^{\xi}P_{j_{20}t}{}^{\xi}-(N_t-m_t)P_{j_{10}t}{}^{\xi}P_{j_{20}t}{}^{\xi}\right] = 0$$

It can be noticed that the estimator of the total of cases with combination (j_1, j_2) of variables y_{i1} and y_{i2} is an unbiased estimator in the case where the three of missing combinations mentioned before are present in the data and only one tree (the joint tree) is used for imputing all the missing records.

5.11.2. Highest Probability Imputation Case

Case 1

Since the missing information is present only for the combination 3 and we are interested in an estimator of the total of cases with category j_1 of y_{i1} and category j_2 of y_{i2}, we use the composite variable y_i^c to estimate that total. In order to assess if this estimator is unbiased, we have to calculate

$$E_{\xi I}\left[\hat{g}_{j^c}-g_{j^c}\right]=E_{\xi I}\left[\sum_{t}\left[\sum_{i=1}^{m_t} I(y_i^c=j^c)+\sum_{i=m_t+1}^{N_t} I(\hat{y}_i^c=j^c)\right]-\sum_{t}\sum_{i=1}^{N_t} I(y_i^c=j^c)\right]$$

As in the Probability Distribution case, this case can be treated as univariate case where the missing variable is a single variable y_i^c with categories $\left\{1,2,\ldots,j^c,\ldots,J^c\right\}$. Therefore, the results are the same as in the univariate case (see *Section 3.6.3* in *Chapter 3*).

Case 2

In this second case, we can be in the presence of Combination 1, Combination 2 and Combination 3 together, that is, variables y_{i1} and y_{i2} can be missing at the same time and also individually.

As before, we want to calculate the expected value of the difference between the estimator and the parameter. That is,

$$E_{\xi I}\left[\hat{g}_{j^c}-g_{j^c}\right] = E_{\xi I}\left[\sum_t\left[\sum_{i=1}^{m_t}I(y_i^c=j^c)+\sum_{i=m_t+1}^{N_t}I(\hat{y}_i^c=j^c)\right]-\sum_t\sum_{i=1}^{N_t}I(y_i^c=j^c)\right]$$

$$= E_\xi\left[\sum_t\left[\sum_{i=1}^{m_t}I(\bar{y}_i=(y_{i1},y_{i2}))+\sum_{i=m_t+1}^{a_t}I(\hat{\bar{y}}_i=(\hat{y}_{i1},y_{i2}))+\sum_{i=a_t+1}^{b_t}I(\hat{\bar{y}}_i=(y_{i1},\hat{y}_{i2}))\right]\right]$$

$$+E_\xi\left[\sum_t\left[\sum_{i=b_t+1}^{N_t}I(\hat{\bar{y}}_i=(\hat{y}_{i1},\hat{y}_{i2}))-\sum_{i=1}^{N_t}I(\bar{y}_i=(y_{i1},y_{i2}))\right]\right]$$

$$=E_\xi\left[\sum_t\left[\sum_{i=m_t+1}^{a_t}I(\hat{\bar{y}}_i=(\hat{y}_{i1},y_{i2}))+\sum_{i=a_t+1}^{b_t}I(\hat{\bar{y}}_i=(y_{i1},\hat{y}_{i2}))+\sum_{i=b_t+1}^{N_t}I(\hat{\bar{y}}_i=(\hat{y}_{i1},\hat{y}_{i2}))\right]\right]$$

$$-E_\xi\left[\sum_t\left[\sum_{i=m_t+1}^{N_t}I(\bar{y}_i=(y_{i1},y_{i2}))\right]\right]$$

$$=E_\xi\left[\sum_t\left[\sum_{i=m_t+1}^{a_t}I(j^*_{1t/j_{20}}=j_{10})I(j_{i2}=j_{20})+\sum_{i=a_t+1}^{b_t}I(j_{i1}=j_{10})I(j^*_{2t/j_{10}}=j_{20})\right]\right]$$

$$+E_\xi\left[\sum_t\left[\sum_{i=b_t+1}^{N_t}I(j^*_{1t/j_{20}}=j_{10})I(j^*_{2t/j_{10}}=j_{20})-\sum_{i=m_t+1}^{N_t}I(j_{i1}=j_{10})I(j_{i2}=j_{20})\right]\right]$$

$$=E_\xi\left[\sum_t\left[I(j^*_{1t/j_{20}}=j_{10})\sum_{i=m_t+1}^{a_t}I(j_{i2}=j_{20})+I(j^*_{2t/j_{10}}=j_{20})\sum_{i=a_t+1}^{b_t}I(j_{i1}=j_{10})\right]\right]$$

$$+E_\xi\left[\sum_t\left[(N_t-b_t)I(j^*_{1t/j_{20}}=j_{10})I(j^*_{2t/j_{10}}=j_{20})\right]\right]$$

$$-E_\xi\left[\sum_t\left[\sum_{i=m_t+1}^{N_t} I(j_{i1}=j_{10})I(j_{i2}=j_{20})\right]\right]$$

$$=\sum_t\left[E_\xi\left[I(j^*_{1t/j_{20}}=j_{10})\sum_{i=m_t+1}^{a_t} I(j_{i2}=j_{20})\right]+E_\xi\left[I(j^*_{2t/j_{10}}=j_{20})\sum_{i=a_t+1}^{b_t} I(j_{i1}=j_{10})\right]\right]$$

$$+\left[\sum_t\left[(N_t-b_t)E_\xi\left[I(j^*_{1t/j_{20}}=j_{10})I(j^*_{2t/j_{10}}=j_{20})\right]\right]\right]$$

$$-\left[\sum_t E_\xi\left[\sum_{i=m_t+1}^{N_t}\left[I(j_{i1}=j_{10})I(j_{i2}=j_{20})\right]\right]\right]$$

$$=\sum_t\left[E_\xi\left[\sum_{i=m_t+1}^{a_t} I(j_{i2}=j_{20})\right]E_\xi\left[I(j^*_{1t/j_{20}}=j_{10})\right]\right]$$

$$+\sum_t\left[E_\xi\left[\sum_{i=a_t+1}^{b_t} I(j_{i1}=j_{10})\right]E_\xi\left[I(j^*_{2t/j_{10}}=j_{20})\right]\right]$$

$$+\left[\sum_t\left[(N_t-b_t)E_\xi\left[I(j^*_{1t/j_{20}}=j_{10})I(j^*_{2t/j_{10}}=j_{20})\right]\right]\right]$$

$$-\left[\sum_t E_\xi\left[\sum_{i=m_t+1}^{N_t}\left[I(j_{i1}=j_{10})I(j_{i2}=j_{20})\right]\right]\right]$$

assuming independence between units,

$$=\sum_t\left[(a_t-m_t)P_{j_{20}}{}^{\xi}P^{\xi}(j^*_{1t}=j_{10})+(b_t-a_t)P_{j_{10}}{}^{\xi}P^{\xi}(j^*_{2t}=j_{20})\right]$$

$$+\sum_t\left[(N_t-b_t)P^{\xi}(j^*_{1t}=j_{10})P^{\xi}(j^*_{2t}=j_{20})-(N_t-m_t)P_{j_{20}}{}^{\xi}P_{j_{10}}{}^{\xi}\right]$$

Given the results for the assessment of bias in the univariate case (i.e. the estimator of the total of cases in category j is not an unbiased estimator), there is not reason why we would think that the quantity presented above as the estimation of the bias in the multivariate case is zero. Therefore, we can say that the estimator of the total of cases with $y_{i1}=j_1$ and $y_{i2}=j_2$ is not an unbiased estimator in the case where the three missing combinations mentioned in *Table 5.6.3* are present in the data and only one tree (the joint tree for y_{i1} *and* y_{i2} together) is used for imputing all the missing records. In this case, the bias will depend on probability of the modal category in each specific terminal node given by the model and also on the number of records missing within each category.

5.11.3. Nearest Neighbour Imputation Case

Case 1

Since this case correspond to Combination 3 only, for the total of cases with category j^c of the variable y_i^c we have

$$E_{\xi I}\left[\hat{g}_j - g_j\right] = E_{\xi I}\left[\sum_t \left[\sum_{i=1}^{m_t} I(y_i = j) + \sum_{i=m_t+1}^{N_t} I(\hat{y}_i = j)\right] - \sum_t \sum_{i=1}^{N_t} I(y_i = j)\right]$$

Therefore, we are again in presence of a univariate case, which was considered in *Section 3.6.4* in *Chapter 3.*

Case 2

Since this case includes y_{i1} and y_{i2} missing at the same time and also individually, we have

$$E_{\xi}\left[\hat{g}_j - g_j\right] = E_{\xi}\left[\sum_t \left[\sum_{i=1}^{m_t} I(y_i = j) + \sum_{i=m_t+1}^{N_t} I(\hat{y}_i = j)\right] - \sum_t \sum_{i=1}^{N_t} I(y_i = j)\right]$$

$$= E_{\xi}\left[\sum_t \left[\sum_{i=1}^{m_t} I(\bar{y}_i = (y_{i1}, y_{i2})) + \sum_{i=m_t+1}^{a_t} I(\hat{\bar{y}}_i = (\hat{y}_{i1}, y_{i2})) + \sum_{i=a_t+1}^{b_t} I(\hat{\bar{y}}_i = (y_{i1}, \hat{y}_{i2}))\right]\right]$$

$$+ E_{\xi}\left[\sum_t \left[\sum_{i=b_t+1}^{N_t} I(\hat{\bar{y}}_i = (\hat{y}_{i1}, \hat{y}_{i2})) - \sum_{i=1}^{N_t} I(\bar{y}_i = (y_{i1}, y_{i2}))\right]\right]$$

$$= E_{\xi}\left[\sum_t \left[\sum_{i=m_t+1}^{a_t} I(\hat{\bar{y}}_i = (\hat{y}_{i1}, y_{i2})) + \sum_{i=a_t+1}^{b_t} I(\hat{\bar{y}}_i = (y_{i1}, \hat{y}_{i2})) + \sum_{i=b_t+1}^{N_t} I(\hat{\bar{y}}_i = (\hat{y}_{i1}, \hat{y}_{i2}))\right]\right]$$

$$- E_{\xi}\left[\sum_t \left[\sum_{i=m_t+1}^{N_t} I(\bar{y}_i = (y_{i1}, y_{i2}))\right]\right]$$

$$= E_{\xi}\left[\sum_t \left[\sum_{i=m_t+1}^{a_t} I(\hat{j}_{i1} = j_{10}) I(j_{i2} = j_{20}) + \sum_{i=a_t+1}^{b_t} I(j_{i1} = j_{10}) I(\hat{j}_{i2} = j_{20})\right]\right]$$

$$+ E_{\xi}\left[\sum_t \left[\sum_{i=b_t+1}^{N_t} I(\hat{j}_{i1} = j_{10}) I(\hat{j}_{i2} = j_{20}) - \sum_{i=m_t+1}^{N_t} I(\bar{y}_i = (j_{10}, j_{20}))\right]\right]$$

$$= E_{\xi}\left[\sum_t \left[\sum_{i'=1}^{m_t} A_{i'} I(j_{i'1} = j_{10}) I(j_{i'2} = j_{20}) + \sum_{i'=1}^{m_t} B_{i'} I(j_{i'1} = j_{10}) I(j_{i'2} = j_{20})\right]\right]$$

$$+E_{\xi}\left[\sum_{t}\left[\sum_{i'=1}^{m_t} C_{i'} I(j_{i'1}=j_{10})I(j_{i'2}=j_{20})-\sum_{i=m_t+1}^{N_t} I(j_{i1}=j_{10})I(j_{i2}=j_{20})\right]\right]$$

$$=\sum_{t}\left[\sum_{i'=1}^{m_t} A_{i'}E_{\xi}\left[I(j_{i'1}=j_{10})I(j_{i'2}=j_{20})\right]+\sum_{i'=1}^{m_t} B_{i'}E_{\xi}\left[I(j_{i'1}=j_{10})I(j_{i'2}=j_{20})\right]\right]$$

$$+\sum_{t}\left[\sum_{i'=1}^{m_t} C_{i'}E_{\xi}\left[I(j_{i'1}=j_{10})I(j_{i'2}=j_{20})\right]-\sum_{i=m_t+1}^{N_t} E_{\xi}\left[I(j_{i1}=j_{10})I(j_{i2}=j_{20})\right]\right]$$

given the definition of $A_{i'}$, $B_{i'}$ and $C_{i'}$, number of times unit i is used as specified below,

$$A_{i'}=\sum_{i\in(m_t+1,a_t)} I(d_{i'i}\le d_{i'l}\ for\ all\ l\in(1,m_t)).$$

$$B_{i'}=\sum_{i\in(a_t+1,b_t)} I(d_{i'i}\le d_{i'l}\ for\ all\ l\in(1,m_t)).$$

$$C_{i'}=\sum_{i\in(b_t+1,N_t)} I(d_{i'i}\le d_{i'l}\ for\ all\ l\in(1,m_t)).$$

and since each donor is used only for imputing cases within the same terminal node

$$=\sum_{t}\left[(a_t-m_t)P_{j_{10}t}{}^{\xi}P_{j_{20}t}{}^{\xi}+(b_t-a_t)P_{j_{20}t}{}^{\xi}P_{j_{10}t}{}^{\xi}+(N_t-b_t)P_{j_{10}t}{}^{\xi}P_{j_{20}t}{}^{\xi}-(N_t-m_t)P_{j_{10}t}{}^{\xi}P_{j_{20}t}{}^{\xi}\right]=0$$

It can be noticed that in this case, the estimator of the total of cases with $y_{i1}=j_1$ and $y_{i2}=j_2$ is an unbiased estimator even when only the joint tree for (y_{i1},y_{i2}) is used for imputing the three different combinations of missing information presented in *Table 5.6.3.*

CHAPTER 6

MULTIVARIATE CASE SIMULATION

6.1 INTRODUCTION

As in *Chapter 4*, the aim of this chapter is to describe the simulation procedure followed when evaluating the imputation performance given the use of classification trees but imputing more than one variable at the same time. Several simulations were carried out using the same database used in the univariate case, which contains synthetic missing values. As well as in the univariate case, in the multivariate simulations different classification trees and imputation methods were used in order to compare the effect of these on the final results. Additionally, different ways of evaluation were applied in order to compare the different procedures.

Moreover, biases and variances were estimated in order to evaluate the properties of the estimators used.

6.2 SIMULATION PROCEDURE

The simulation procedure carried out for the multivariate case is basically the same procedure employed in the univariate case but imputing more than one variable at the same time. Therefore, since most of the features were described in *Chapter 4*, a review of the most important aspects and description of new ones will be shown in this section.

1. ***Generation of the synthetic database.*** The database used in the multivariate case is the same used in the univariate case.

2. ***Growing trees.*** As in the univariate case, different trees were grown for each target variable, but in this case using composite variables describe in *Section 5.5* in *Chapter 5.*

    2.1 For each target variable (composite variables), three different tree-sizes were used in the analysis in order to compare the effect of the size of the tree on the imputation results. The selection of the sizes was the same procedure explained in *Section 4.5.2* in the simulation chapter for the univariate case.

    2.2 After all the trees had been grown, the records with missing values in the target variables were dropped into each tree to find out which terminal node they will end up in order to carry out with the imputation. This procedure was followed for the different tree-sizes.

3. ***Imputing.*** After the different trees were generated, imputation was carried out independently for each of the trees for the composite variables used.

    3.1 The three different imputation methods were combined with the three different tree sizes to obtain 9 different imputation results for each target variable. This was made using trees grown with the complete database.

    3.2 For each of the trees, the imputation was carried out independently into each terminal node. Then, the results were summarised in order to compare them with the results from other trees.

4. ***Evaluation*** Different graphs, tests, biases and variances were used for evaluation of the imputation.

    4.1 Cross-tabulations between the imputed values and the real values were obtained for all of the possible combinations of tree sizes and imputation methods.

    4.2 Graphs were created for any of the above tables in order to compare preservation of individual marginal and joint distributions and preservation of individual values.

4.3 Tests were also used for each of the cross-tabulations in order to confirm the preservation of individual marginal and joint distributions and preservation of individual values.

4.4 Biases and variances were estimated for most of the composite variables imputed in order to assess the properties of the estimators used.

6.3 DATA

6.3.1 Data Description

The database used for the analysis of the multivariate case consists basically of the same database used for the univariate case (see *Section 4.3.1*). The only difference with respect to the univariate case is the variables used as target variables in the analysis.

In the multivariate case, many variables can be missing at the same time. Therefore, the target variables used in these multivariate simulations are basically combinations of two or more single variables.

Because the imputation process in this work requires the generation of a tree for the target variable as a first step, and since those trees are grown for single variables, composite variables were created.

As explained in *Section 5.5* in *Chapter 5*, a composite variable is defined by the cross-classification of two or more single variables with categories defined by the combination of the categories of each of the variables involved.

Since all variables on the database are categorical, combinations of these variables also correspond to categorical variables.

A list of the composite variables used in this thesis for the multivariate analysis is shown in *Table 6.3.1*, including a description of the combinations and the definition of their new categories.

Tables 6.3.1

Composite Variable Definitions

Table A
Country of Birth - Ethnic (COB - ETHNIC)

Cob	Ethnic	Cob	Ethnic	New Code
1	1	UK	White	1
1	2	UK	Any black including mixed	2
1	3	UK	Asian	3
1	4	UK	China / Other including other mixed	4
2	1	Europe / USA	White	5
2	2	Europe / USA	Any black including mixed	6
2	3	Europe / USA	Asian	7
2	4	Europe / USA	China / Other including other mixed	8
3	1	Indian Sub-continent	White	9
3	2	Indian Sub-continent	Any black including mixed	10
3	3	Indian Sub-continent	Asian	11
3	4	Indian Sub-continent	China / Other including other mixed	12
4	1	Africa / Caribbean	White	13
4	2	Africa / Caribbean	Any black including mixed	14
4	3	Africa / Caribbean	Asian	15
4	4	Africa / Caribbean	China / Other including other mixed	16
5	1	Asia / Central and South America / Other	White	17
5	2	Asia / Central and South America / Other	Any black including mixed	18
5	3	Asia / Central and South America / Other	Asian	19
5	4	Asia / Central and South America / Other	China / Other including other mixed	20

Table B
Country of Birth - Long term Illness (COB - LTILL)

Cob	Ltill	Cob	Ltill	New Code
1	1	UK	Has a health problem	1
1	2	UK	Does not have a health problem	2
2	1	Europe / USA	Has a health problem	3
2	2	Europe / USA	Does not have a health problem	4
3	1	Indian Sub-continent	Has a health problem	5
3	2	Indian Sub-continent	Does not have a health problem	6
4	1	Africa / Caribbean	Has a health problem	7
4	2	Africa / Caribbean	Does not have a health problem	8
5	1	Asia / Central and South America / Other	Has a health problem	9
5	2	Asia / Central and South America / Other	Does not have a health problem	10

Table C
Ethnic - Long Term Illness (ETHNIC - LTILL)

Ethnic	Ltill	Ethnic	Ltill	New Code
1	1	White	Has a health problem	1
1	2	White	Does not have a health problem	2
2	1	Any black including mixed	Has a health problem	3
2	2	Any black including mixed	Does not have a health problem	4
3	1	Asian	Has a health problem	5
3	2	Asian	Does not have a health problem	6
4	1	China / Other including other mixed	Has a health problem	7
4	2	China / Other including other mixed	Does not have a health problem	8

Table D
Country of Birth - Ethnic - Long Term Illness (COB - ETHNIC - LTILL)

Cob	Ethnic	Ltill	Cob	Ethnic	Ltill	New Code
1	1	1	UK	White	Has a health problem	1
1	1	2	UK	White	Does not have a health problem	2
1	2	1	UK	Any black including mixed	Has a health problem	3
1	2	2	UK	Any black including mixed	Does not have a health problem	4
1	3	1	UK	Asian	Has a health problem	5
1	3	2	UK	Asian	Does not have a health problem	6
1	4	1	UK	China / Other including other mixed	Has a health problem	7
1	4	2	UK	China / Other including other mixed	Does not have a health problem	8
2	1	1	Europe / USA	White	Has a health problem	9
2	1	2	Europe / USA	White	Does not have a health problem	10
2	2	1	Europe / USA	Any black including mixed	Has a health problem	11
2	2	2	Europe / USA	Any black including mixed	Does not have a health problem	12
2	3	1	Europe / USA	Asian	Has a health problem	13
2	3	2	Europe / USA	Asian	Does not have a health problem	14
2	4	1	Europe / USA	China / Other including other mixed	Has a health problem	15
2	4	2	Europe / USA	China / Other including other mixed	Does not have a health problem	16
3	1	1	Indian Sub-continent	White	Has a health problem	17
3	1	2	Indian Sub-continent	White	Does not have a health problem	18
3	2	1	Indian Sub-continent	Any black including mixed	Has a health problem	19
3	2	2	Indian Sub-continent	Any black including mixed	Does not have a health problem	20
3	3	1	Indian Sub-continent	Asian	Has a health problem	21
3	3	2	Indian Sub-continent	Asian	Does not have a health problem	22
3	4	1	Indian Sub-continent	China / Other including other mixed	Has a health problem	23
3	4	2	Indian Sub-continent	China / Other including other mixed	Does not have a health problem	24
4	1	1	Africa / Caribbean	White	Has a health problem	25
4	1	2	Africa / Caribbean	White	Does not have a health problem	26
4	2	1	Africa / Caribbean	Any black including mixed	Has a health problem	27
4	2	2	Africa / Caribbean	Any black including mixed	Does not have a health problem	28
4	3	1	Africa / Caribbean	Asian	Has a health problem	29
4	3	2	Africa / Caribbean	Asian	Does not have a health problem	30
4	4	1	Africa / Caribbean	China / Other including other mixed	Has a health problem	31
4	4	2	Africa / Caribbean	China / Other including other mixed	Does not have a health problem	32
5	1	1	Asia / Central and South America / Other	White	Has a health problem	33
5	1	2	Asia / Central and South America / Other	White	Does not have a health problem	34
5	2	1	Asia / Central and South America / Other	Any black including mixed	Has a health problem	35
5	2	2	Asia / Central and South America / Other	Any black including mixed	Does not have a health problem	36
5	3	1	Asia / Central and South America / Other	Asian	Has a health problem	37
5	3	2	Asia / Central and South America / Other	Asian	Does not have a health problem	38
5	4	1	Asia / Central and South America / Other	China / Other including other mixed	Has a health problem	39
5	4	2	Asia / Central and South America / Other	China / Other including other mixed	Does not have a health problem	40

6.3.2 Pattern of missing information

The second stage involved finding the pattern of missing information present in the data in order to create artificial holes for evaluating the imputation process. Since the database used in this chapter is the same used in the univariate case, the pattern of missing information is still valid as it includes all the possible combinations of missing information, i.e. two, three and more variables missing at the same time.

As mentioned in *Section 4.3.2* in *Chapter 4*, the size of the database used (original database) is 222872 records with 23 variables. The total number of records with missing information is 24116, which represents 10.82 % of the original database (222872 records).

Table 6.3.2.1 includes the combinations of variables, the total numbers of missing cases and percentages of missingness used in the multivariate simulations. The complete list of combinations of missing variable with their respective percentages can be see in *Appendix 1.*

Table 6.3.2.1

Combinations and percentages of missing information used for the simulations in the multivariate case

Variable				
Country of Birth	Ethnic	Long Term Illness	Total	Perc.
	■	■	465	0.74
■		■	225	1.04
■	■		159	2.16
■	■	■	464	2.16

6.3.3. Databases used in the analysis

Table 6.3.3.1 shows the different sizes of the databases (original, complete and synthetic) and percentages of missing information as also explained in *Chapter 4.*

Table 6.3.3.1

Databases sizes and Percentages of missing information

Database	Size	Complete Information	Missing Information
Original Database	222872	198756	10.820%
Complete Database	198756	198756	None
Synthetic Database	198756	177236	10.827%

It is important to point out that the analysis could be done including missing information for the covariates, however, for simplicity, only complete information is included in the generation of the tree. Additionally, a previous study by Mesa, Tsai and Chambers (2000) shows that the inclusion of missing information for growing the tree seems to have no impact on the results when using the same imputation procedures used in this thesis.

Every time a combination of variables with missing information is chosen to define a target variable, the remaining information also changes since different variables are left as covariates (auxiliary variables). Therefore, depending on the combination used as a target, the databases used for the analysis (growing trees, etc.) are different.

The sizes of the four databases used, depending on the target combination studied, are shown in the next table

Table 6.3.3.2

Databases sizes for the multivariate case

Variable	Database Size	Missing Information
Any	198756 (complete records)	None
Ethnic - Ltill	198756 - 465 = 198291 (Records with Ethnic and Ltill missing)	Rest of the variable
Cob - Ltill	198756 - 225 = 198531 (Records with Cob and Ltill missing)	Rest of the variable
Ethnic - Cob	198756 - 159 = 198597 (Records with Ethnic and Cob missing)	Rest of the variable
Ethnic - Cob - Ltill	198756 - 464 = 198292 (Records with Ethnic, Ltill and Cob missing)	Rest of the variable

It is important to point out that for simplicity, only one database is used for growing the tree independently of the missing combination studied. The database used is the one containing only fully observed records for all the variables and it is the "complete database" shown in *Table 6.3.3.1.*

6.4 CLASSIFICATION

Once the database is ready, the first step of the process is the classification of the units into terminal nodes using CART. In this case, all features related to classification are exactly the same as used in the univariate case with the difference that the target variable is now a composite variable especially created for this task.

6.4.1 Splitting Criterion

The composite variable used as a target variable is basically treated as a single categorical variable as it is a product of the combination of two or more single variables, therefore, the procedure followed for generating the tree was exactly the same as in the univariate case. That is, the splitting criterion used consists on an impurity function defined by the Gini index. Again, costs for misclassifying any class j_1 as a class j_2 are taken equal to 1 for all $j_1 \neq j_2$.

6.4.2 Class Assignment Rule

In the multivariate case, when a composite variable is classified, two or more variables are classified at the same time. Thus, each terminal node is assigned a specific category of the composite variable, which is a combination of categories of two or more variables. For example, if a terminal node has been assigned category 5 of the composite variable Ethnic-Long term illness in *Table 6.3.1 C* in *Section 6.3.1*, that means, the individual variables Ethnic and Long term illness are assigned categories 3 and category 1 respectively. The class assignment rule used in the multivariate case is the same used in the univariate case, plurality rule, see *Section 4.4.2.*

6.4.3 Surrogates

As mentioned in *Chapters 2* and *5*, surrogates could be very useful tools for classification for imputation. They can be used for classifying elements with missing information in the auxiliary variables, allowing for the use of as much information as possible; and also, they can be used for imputing in a sequential way.
The use of surrogates for imputation is explained in *Section 5.7* in *Chapter 5*, however, simulations for this aspect are not carried out due to time constraints.

6.5. TREE PROCESS

6.5.1 Growing the tree

Once the variable to be imputed is selected, a tree for that variable is generated. The process of growing a tree is as explained in *Section 4.5.1* in the univariate case. The only difference is that the target variable is now a composite variable, which includes two or more single variables.

6.5.2. Selection of the tree size

The tree size selection was carried out following the same procedure explained in *Section 4.5.2* in *Chapter 4*, finding a kind of compromise between misclassification rate and number of terminal nodes. Thus, three different tree-sizes were selected for each target variable (small, medium and large). These sizes depend on the output obtained for each variable from the software. In many cases, the optimal tree given by CART was included as one of the trees used for that target variable. In the cases where the optimal tree was too large (with too many terminal nodes), this was not included because of processing time. The tree sizes for the different target variables used in the multivariate case are as follows:

Table 6.5.2.1

Tree sizes for the multivariate case

Target variable	Name	Size of the trees (number of terminal nodes)
COB-ETHNIC	COBETH	10 - 18 (optimal)
COB-LTILL	COBLTI	8 - 15 - 28
ETHNIC-LTILL	ETHLTI	4 - 15 - 27
COB-ETHNIC-LTILL	COETLT	5 - 12 - 23

Since the optimal tree given by CART is considered to be "the optimal" in terms of complexity and misclassification rates, trees larger than the optimal size were not used.

6.5.3. Classifying the records for imputation

After the tree is generated, the records with missing information in the response variable were dropped down the tree in order to identify the terminal nodes (imputation classes) in which those records end up. This procedure is explained in *Section 4.5.3.*

6.6. IMPUTATION METHODS

Once the classification tree is constructed, each imputation method is applied independently within each terminal node of the tree. The final imputation results are evaluated for the tree as a whole by totalling the imputation results obtained at each terminal node.

The three different imputation methods used in the univariate case were also used in these simulations for the multivariate case. These are Frequency Distribution method, Highest Probability method and Nearest Neighbour method and they are recalled in *Section 5.9* in *Chapter 5* as well.

6.7. EVALUATION OF THE IMPUTATION PERFORMANCE

As described in *Section 4.7* in *Chapter 4,* the imputation procedure can be evaluated from different perspectives. Since in this case the imputation is done for more than one variable at the same time, the evaluation of the performance of the imputation must take into account the following aspects,

- ✓ A comparison of the joint distributions for the combination of variables
- ✓ A comparison of the individual values
- ✓ Assessment of the properties of the estimator used

In general, different comparisons can be done depending on the area to be evaluated.

1. To assess the impact of using a classification tree for imputation, comparisons of the results of imputation using trees and not using trees can be done.
2. To evaluate the performance of the different imputation methods when using classification trees, comparisons between the results obtained using different imputation methods can be done.
3. To evaluate the properties of the estimators used in the analysis, bias and variance can be estimated.
4. In addition, if more details want to be given, comparisons can be made between the different categories of the variable being imputed.

It is important to point out that in this thesis, the main aspect to be analysed is the differences in the imputation performance regarding the use of classification trees for forming the imputation classes.

In order to evaluate these aspects, three different methods were used:

- ✓ Graphical comparison
- ✓ Test of agreement
- ✓ Biases and variances

6.7.1. Graphical Comparison

Both kind of graphs, for comparing joint distributions and individual values, used in the univariate case (see *Section 4.7.2*) were also employed in the multivariate case. Then, a group of graphs, for each combination between the different tree sizes and the different imputation methods were obtained in the multivariate case. They were used to compare the imputation performance when imputing more than one variable at the time using different tree-sizes and different imputation methods.

Cross-tabulations between the real and the imputed values were previously obtained for producing the graphs.

6.7.2. Test of agreement

In order to confirm the results obtained from the graphs described in the last section, both statistics Wald statistics and Diagonal statistics described in *Section 4.7.3* were also employed in the multivariate case. The Wald statistics was used for comparing joint distributions while the Diagonal statistics for comparing individual values.

The outcomes are presented in the results section.

6.7.3.Biases and variances

In order to assess the properties of the estimators used, bias and variance were estimated. As mentioned in *Chapter 5*, in this case, estimators of the variance were not obtained (not even in theory).

Simulation for the biases and variances

The use of simulation to estimate the biases and variances in the multivariate case were carried out in the same way as the simulations explained in the univariate case (see *Section 4.7.4* for more details).

1. ***Generation of the databases.*** The databases used in the multivariate case were the same databases (sample databases) generated in the univariate case. The 1000 sample databases created in the univariate case were big enough to include the multivariate missing cases studied in this thesis.

2. ***Generation of the synthetic holes.*** As in the univariate case, the patterns of missing information for the multivariate missing case studied in this thesis were replicated on the 1000 sample databases at random as explained in *Section 4.7.4* in *Chapter 4*. The composite variables used and their missing percentages are shown in *Table 6.3.2.1*.

3. ***Classifying the records for imputation.*** After a classification tree was created for each composite variable used in the study, each database was divided in the corresponding number of terminal nodes depending on the size of the tree used. Records with missing information for the target variable were classified in order to generate the pool of recipients to carry out the imputation procedure. As in the univariate case, since the size of the tree does not seem to have a major impact on the imputation results, only one tree size was used for the simulations of biases and variances described in this chapter.

4. ***Imputation.*** After having the set of donors generated by the tree and recipients generated in point 3 before for each database classified into two different groups, the imputation procedures were applied in order to obtain estimates for the biases and variances. The results for each group (terminal nodes) were totalled and comparisons for trees as wholes were obtained. Then, 1000 estimates were calculated (each for each sample database) in order to measure biases and variances.

It is important to point out that since Nearest Neighbour procedure is very much time consuming, for this part of the simulations only two imputation methods, Frequency Distribution and Highest Probability, were used.

5. ***Calculation of the biases and variances.*** Once the 1000 imputations were obtained for the two different imputation methods used in the multivariate case, biases and variances as described in *Section 4.7.4* were obtained. Results and the correspondent comments are presented in the results.

6.8.RESULTS

This section introduces the most notable findings obtained from the analysis regarding the use of tree-based models for imputation in the multivariate case. As in *Chapter 4,* we divide it in different sections as follows

6.8.1 Using trees

Table 6.8.1.1 shows the values of the Wald statistic for the different target variables, different imputation methods and different tree sizes

Table 6.8.1.1

Wald Statistic and P-values for the multivariate case

			Wald Statistic			P-values		
Variable	Tree Size	d.f.	Freq. Dist.	High. Prob.	Near. Neig.	Freq. Dist.	High. Prob.	Near. Neig.
cobeth	10	19	3.49	26.00	4.00	0.99	0.13	0.99
	18	19	3.54	26.00	2.83	0.99	0.13	0.99
	No Tree	19	4.07	26.00	5.18	0.99	0.13	0.99
coblti	8	9	4.01	52.75	2.78	0.91	0.00	0.97
	15	9	6.86	50.75	8.15	0.65	0.00	0.51
	28	9	7.90	48.32	8.58	0.54	0.00	0.47
	No Tree	9	2.17	73.00	10.19	0.98	0.00	0.33
ethlti	4	7	3.43	108.28	10.16	0.84	0.00	0.17
	15	7	4.66	86.52	10.55	0.70	0.00	0.15
	27	7	7.80	87.66	3.83	0.34	0.00	0.79
	No Tree	7	2.58	170.00	8.04	0.92	0.00	0.32
coetlt	5	39	24.98	184.90	29.58	0.96	0.00	0.86
	12	39	23.08	183.90	32.70	0.97	0.00	0.75
	23	39	26.40	181.56	24.44	0.93	0.00	0.96
	No Tree	39	17.87	193.00	25.03	0.99	0.00	0.95

As explained in *Section 4.8.1*, small values of the Wald statistics (or equivalently, big values for the p-value) suggest no evidence to reject the hypothesis that marginal distributions are maintained and vice versa. Since the degree of freedom for each variable varies depending on the number of categories (i.e. each variable has a different critical value), we simplify the analysis by using p-values.

In terms of preservation of marginal distribution (joint marginal distributions in this case), it can be seen from this table there is not major impact when using trees. The values of the Wald statistics (p-values) indicate that the marginal distributions are maintained even when trees are not used, except for the cases in the Highest Probability method. That is, in the Highest Probability method, none of the distributions is preserved in any of the cases (using or not trees). However, it can be said that the use of trees in the case of Highest Probability method improves the performance of the imputation in term of preservation of distributions, as it will be explained in the next example.
Graphical representation of comparisons between real and imputed distributions can be seen in *Appendix 4*.

In the next example, *Tables 6.8.1.2* show a cross tabulation between real and imputed values for the variable Ethnic - Long term illness using Highest Probability imputation method, two different tree sizes and no tree. In all the tables, the rows represent the values of the imputed variable and the columns represent the real values of the variable. In this case, 465 records were imputed

Tables 6.8.1.2

Cross-tabulation between Real and Imputed Values for the variable Ethnic-Ltill

Table A
4 Terminal Nodes

	Real								
Imputed	1	2	3	4	5	6	7	8	Total
2	45	289	0	48	1	4	0	12	399
4	1	2	6	37	0	2	2	0	50
6	0	0	0	0	0	4	0	0	4
8	0	4	0	4	0	0	0	4	12
Total	46	295	6	89	1	10	2	16	465

Table B
15 Terminal Nodes

	Real								
Imputed	1	2	3	4	5	6	7	8	Total
1	5	4	0	0	0	0	0	0	9
2	40	285	0	48	1	4	0	12	390
3	0	0	2	3	0	0	0	0	5
4	1	2	4	34	0	2	2	0	45
6	0	0	0	0	0	4	0	0	4
8	0	4	0	4	0	0	0	4	12
Total	46	295	6	89	1	10	2	16	465

Table C
No Trees

	Real								
Imputed	1	2	3	4	5	6	7	8	Total
2	46	295	6	89	1	10	2	16	465
Total	46	295	6	89	1	10	2	16	465

From *Table 6.8.1.2 C* we have that this variable contains eight different categories. Most of them are used for imputation in the case where a classification tree is used *Table 6.8.1.2 A* and *Table 6.8.1.2 B*, depending on which terminal node each recipient ends up. In this example, the trees used have 4 and 15 terminal nodes, which even when it has very few number of terminal nodes for the first one, allows for the use of many categories for imputing. However, in the case where a classification tree is not used (*Table 6.8.1.2 C*), only the category with highest probability was used for imputing, which is category 2.

This example shows that employing a classification tree ensure the use of most of the categories of the variable for imputation, even when the tree does not have a large number of terminal nodes. However, when trees are not used, only one category is used for imputation. That is because the imputation when using trees is made at each terminal node and categories for imputing will depend on the class assignment that define the terminal nodes class, while when trees are not used, the imputation will be made employing the category with highest probability in the whole database, which will be just one.

Therefore, we can say that even when none of the distributions are preserved for the Highest Probability method, in this case, the use of the tree improve the distributions obtained after imputation. This aspect can also be seen in a graphical way in *Appendix 4* and *Appendix 5*.

There is an important aspect to point out in this analysis in terms of preservation of marginal distributions. In most of the cases (almost all of them) the marginal individual distribution in the case of single variables and joint distributions in the case of composite variables are maintained. A previous analysis made by Mesa, Tsai and Chambers (2000) shows that when imputation is done for a composite variable using this procedure not only joint distributions are preserved but the individual marginal distributions of the single variables involved in that composite variable are also preserved. The work made by Mesa, Tsai and Chambers presents an example when two variables, Primary activity last week (ALWPRIM) and Long term illness (LTILL) were imputed at the same time. In this case, the joint distribution Primary activity last week - Long term illness was maintained, as well as the individual marginal distributions for Primary activity last week and Long term illness separately.

This is a really valuable achievement since it is important to uphold relationships between variables without loosing the shape of the distribution of the individual variables.

Similar simulations have not been done for the variables used in this analysis because of time constraints.

Table 6.8.1.3 show the results of the Diagonal Statistic for the multivariate case.

Table 6.8.1.3

Diagonal Statistic (z_D) for the multivariate case

		Diagonal Statistic		
Variable	Tree Size	Freq. Dist.	High. Prob.	Near. Neig.
cobeth	10	5.26	2.25	5.393
	18	4.10	2.25	3.349
	No Tree	5.90	2.25	5.02
coblti	8	8.33	4.57	7.63
	15	7.74	4.39	6.96
	28	8.09	4.39	7.51
	No Tree	11.86	5.92	6.96
ethlti	4	13.87	7.16	12.11
	15	11.86	7.16	11.13
	27	11.78	6.90	11.13
	No Tree	18.50	9.89	11.21
coetlt	5	23.27	11.72	20.51
	12	22.98	11.64	20.00
	23	22.83	11.72	20.00
	No Tree	22.98	11.72	20.90

As explained in *Section 4.7.3* in *Chapter 4*, provided one cannot reject the hypothesis that the imputation method preserves the marginal distribution using the Wald statistic, the preservation of individual values can be tested by using the confidence interval for D (proportion of incorrectly imputed cases). In this case, $D-2\sqrt{V(D)}$ should be less than zero in order to have some evidences that the individual values are preserved. In other words, if $z_D - 2 < 0$, then, the individual values can be said to be preserved, with $z_D = \frac{D}{\sqrt{\hat{V}(D)}}$.

It can be noticed that if a confidence interval is calculated as explained before, $z_D - 2$ is closer to zero when trees are used than in the case of not using trees in many cases. This improvement can be observed from the point of view of percentage of records correctly imputed and it will be explained later.

In the case of Nearest Neighbour method, the values of z_D are very similar in both cases, when using and not trees. Then, it cannot be said the method performs better when using trees than when trees are not used.

A graphical representation of the preservation of individual values is shown in *Appendix 5*.

It is important to point out that even when the values of the diagonal statistic in the case of using tree are also large (compared with not using trees) as shown in *Table 6.8.1.1* (i.e. individual values are not preserved), the percentages of records correctly imputed obtained from these results show an improvement on the imputation performance with respect to the results of the cases where trees are not used in many cases (or at least remain equal). This can be seen in the next table.

Table 6.8.1.4 presents the "improvement" for the different combinations between tree sizes and imputation methods for the different variables. This measure of improvement is based on the percentage of records correctly imputed when trees are not used and their differences with the percentage of records correctly imputed when trees are used as explained in *Section 4.8.3* in *Chapter 4*. See *Appendix 7* for graphical representation.

If a comparison between the results from the case where trees are used and that where trees are not used is made, we will notice that there is always an improvement in terms of records correctly imputed when using trees for the Frequency Distribution except for the variable Cob-Ethnic-Ltill and in many cases for the Highest Probability method. However, in the case of Nearest Neighbour, the improvement seems to be more variable since some of them have more correctly imputed records when using trees and some have more correctly imputed records when trees are not used.

It is clear that the highest improvement is always for Frequency Distribution method and this improvement can reach more than 30% in some cases for the multivariate case.

Table 6.8.1.4

Improvement by variable, tree size and imputation method for the multivariate case

Variable	Tree-size	Freq. Distrib.	High. Prob.	Near. Neigh.
Cobeth	10	5.00	0.00	-2.80
	18	15.00	0.00	14.02
	No Tree	0.00	0.00	0.00
Coblti	8	25.00	9.21	3.82
	15	29.81	10.53	8.40
	28	26.92	10.53	4.58
	No Tree	0.00	0.00	0.00
Ethlti	4	22.28	13.22	-3.96
	15	33.66	13.22	0.36
	27	34.16	14.58	0.36
	No Tree	0.00	0.00	0.00
Coetlt	5	-1.20	0.00	1.65
	12	0.00	0.37	3.85
	23	0.60	0.00	3.85
	No Tree	0.00	0.00	0.00

Table 6.8.1.5 shows the percentage of records incorrectly imputed by variable, tree size and imputation method, the misclassification rate by tree size and variables as well as the percentage of missing information by variable

Table 6.8.1.5

Percentage of missing data, percentage of records incorrectly imputed and misclassification rate by variable, imputation method and tree size for the multivariate case.

Variable	Percentage of missing data	Tree Size	Probability Distribution	Highest Probability	Nearest Neighbour	Misclassification Rate
Cob_eth	0.74	10	33.96	16.35	34.59	36.41
		18	27.67	16.35	23.27	36.40
		No tree	37.10	16.35	32.70	------
Cob_lti	1.24	8	42.22	26.22	39.55	25.00
		15	40.00	25.33	36.88	24.53
		28	41.33	25.33	39.11	24.29
		No tree	53.77	32.44	41.77	-----
Eth_lti	2.16	4	46.88	28.17	42.58	29.38
		15	41.93	28.17	40.00	28.80
		27	41.72	27.31	40.00	28.71
		No tree	56.55	36.55	40.21	------
Co_et_lt	2.16	5	64.43	41.59	60.12	44.18
		12	64.00	41.37	59.26	44.14
		23	63.79	41.59	59.26	44.12
		No tree	64.00	41.59	60.77	------

It can be noticed that the relationship between the percentage of records incorrectly imputed and the misclassification rate when using trees is not as obvious as in the univariate case.

Another interesting point to notice from this table is the fact that the lowest percentage of records incorrectly imputed is for the Highest Probability method, even when trees are not used. That is because the majority of the population is always concentrated in one category (or two), which is the one used for imputation by this method. Probability Distribution and Nearest Neighbour present similar percentages of records incorrectly imputed most of the time.

Additionally, we can see that the highest misclassification rates are for the variables with more categories as CO-ET-LT and COB-ETH.

6.8.2 Comparing Tree-Sizes

It can be seen from *Table 6.8.1.1* and *Table 6.8.1.3* in the last section that the changes in the Wald statistic and the Diagonal statistic are not big enough to alter the conclusion that the imputation performance is not affected by the size of the tree. Additionally, the changes on both statistics do not follow similar pattern for all of the cases. Sometimes the best results are obtained from the smallest trees and sometimes from the biggest trees or even from the medium size trees. However, since all of the values of the Wald statistics are small enough to not reject the hypothesis that marginal distributions are maintained, and the values of the Diagonal statistics are big enough to reject the hypothesis that individual values are preserved, it can be said that there are not considerable changes on the results when using different sizes of trees.

Therefore, the main conclusion about using different sizes for the tree is that increasing the size does not necessarily improve the imputation performance. That is, using complex trees does not necessary lead to better imputation results.

Moreover, looking at *Tables 6.8.1.1* and *6.8.1.3* it can be noticed that there are no major differences between using 10 or 18 terminal nodes for the variable Country of birth - Ethnic (COB-ETHNIC) in terms of the values of the Wald statistics and Diagonal statistics, even when the tree with 18 terminal nodes is the optimal tree given by CART. Then, we can say that the use of the optimal tree given by CART does not seem to make significant improvement in the performance of the imputation. The optimal tree given by CART is meant to be optimal in terms of complexity and misclassification rate. In this sense, the use of the optimal tree could be expected to give the best performance, however, it can be observed from the results that this hypothesis is not necessarily correct. More results about this aspect, leading to the same conclusions, can be found in Mesa, Tsai and Chambers (2000).

As shown in *Table 6.8.1.5,* there seems to be no relationship between the misclassification rate and the percentage of records correctly imputed for each composite variable. The percentage of records correctly imputed look stable as well as the misclassification rate within each variable for the different tree sizes.

6.8.3 Comparing Imputation Methods

As said in *Section 6.8.1,* it can be seen from *Table 6.8.1.1* that Frequency Distribution and Nearest Neighbour perform very well in term of preservation of marginal distributions given the P-values for the Wald statistics when using trees, which is not the case of Highest Probability method.

In the case of Highest Probability, there is not preservation of marginal distribution in any of the cases, as observed in *Table 6.8.1.1,* however, there is an improvement on the distribution of the imputed values when trees are used as explained in *Section 6.8.1.*

In terms of preservation of individual values we can see that none of the imputation procedures used in this research achieve this aim. However, we can see some differences in the values of the Highest Probability and the rest of the imputation methods. Highest Probability method has always lower values in the diagonal statistics. This fact can also be seen in the next table.

Table 6.8.3.1, as well as *Appendix 6* (in graphical terms), shows the percentages of records correctly imputed when using or not tree for the different imputation methods and different tree sizes.

Table 6.8.3.1

Percentage of Cases Correctly Imputed for the multivariate case

Variable	Tree-size	Freq. Distrb.	High. Prob.	Near. Neigh.
cobeth	10	66.04	83.65	65.41
	18	72.33	83.65	76.73
	No Tree	62.89	83.65	67.30
cotlti	8	57.78	73.78	60.44
	15	60.00	74.67	63.11
	28	58.67	74.67	60.89
	No Tree	46.22	67.56	58.22
ethlti	4	53.12	71.83	57.42
	15	58.06	71.83	60.00
	27	58.28	72.69	60.00
	No Tree	43.44	63.44	59.78
coetlt	5	35.56	58.41	39.87
	12	35.99	58.62	40.73
	23	36.21	58.41	40.73
	No Tree	35.99	58.41	39.22

It can be seen that always, the higher percentage is obtained when using Highest Probability method.

Even for the strangest cases (Country of birth-Ethnic-Long term illness, COETLT, and Country of birth- Ethnic, COBETH), the lowest percentage of records correctly imputed obtained with the Highest Probability method is over 58%.
Thus, we can say that the best method in preserving individual values is the Highest Probability with over 80% of the cases correctly imputed in some situations.
The percentage of records correctly imputed with this method depends, in a way, on the shape of the distribution when using trees and of course on the accuracy of the classification tree.

In the case of Frequency Distribution, there is always an improvement when using tree. This improvement is not evident when comparing marginal distributions but it can be observed when comparing individual values.

Table 6.8.1.5, as well as *Appendix 8,* show the relationship between the percentage of misclassification rate and the percentage of records correctly imputed. It can be seen that the higher the misclassification rate, the lower the number of records correctly imputed. This applies to all the methods with some exceptions in the Nearest Neighbour.
It can also be noticed from *Table 6.8.1.5* that as in the univariate case, the highest percentages of records incorrectly imputed always corresponds to the Frequency Distribution methods, followed by Nearest Neighbour and Highest Probability respectively.

It can be noticed from the Wald statistic values in *Table 6.8.1.1,* Diagonal values in *Table 6.8.1.3* and in the percentage of records correctly imputed in *Table 6.8.3.1* that in the case of Nearest Neighbour, the use of the tree does not have very much impact on the results. The results remain the same when comparing both marginal distributions and individual values. Additionally, the percentage of records correctly imputed remains fairly stable when using Nearest Neighbour. Therefore, we can say that, in general, the use of trees does not make any improvement in the results when using Nearest Neighbour, probably because the nearest neighbour donor will be found either using or not classification. The use of the tree will probably improve the time consumed in the imputation process.

6.8.4 Comparing Categories

Tables 6.8.4.1 contain the percentage of records incorrectly imputed by imputation methods, tree sizes and categories, as well as the misclassification rates obtained from the different tree sizes by categories of the target variables

Tables 6.8.4.1

Misclassification rates by tree sizes and categories and percentage of records incorrectly imputed by imputation method, tree size and categories for the multivariate case

Table A

Variable: ETH_LTILL

					Percentage of Records Incorrectly Imputed											
		Misclass. Rate			Frequency Distribution				Highest Probability				Nearest Neighbour			
Cat.	Records	8 TN	15 TN	27 TN	8 TN	15 TN	27 TN	N TRE	8 TN	15 TN	27 TN	N TRE	8 TN	15 TN	27 TN	N TRE
1	16006	100.00	90.42	89.10	95.65	67.39	67.39	86.96	100.00	89.13	86.96	100.00	67.39	73.91	63.04	71.74
2	110727	2.78	3.41	3.36	30.17	27.46	27.46	38.98	2.03	3.39	2.71	0.00	31.53	22.71	28.47	24.75
3	3352	100.00	81.00	74.67	66.67	66.67	83.33	100.00	100.00	66.67	50.00	100.00	50.00	83.33	50.00	66.67
4	32064	60.81	62.17	62.75	62.92	61.80	61.80	82.02	58.43	61.80	61.80	100.00	53.93	61.80	52.81	60.67
5	677	100.00	100.00	85.67	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
6	5609	60.35	60.35	61.76	80.00	70.00	60.00	100.00	60.00	60.00	60.00	100.00	80.00	80.00	70.00	70.00
7	718	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
8	8083	66.31	66.31	68.29	87.50	87.50	81.25	100.00	75.00	75.00	75.00	100.00	75.00	87.50	81.25	81.25

Table B

Variable: Country of birth - Long term illness

					Percentage of records incorrectly imputed											
		Misclass. Rate			Frequency Distribution				Highest Probability				Nearest Neighbour			
Cat.	Records	8 TN	15 TN	28 TN	8 TN	15 TN	28 TN	N TRE	8 TN	15 TN	28 TN	N TRE	8 TN	15 TN	28 TN	N TRE
1	15687	100.00	89.59	89.59	100.00	70.00	75.00	95.00	100.00	90.00	90.00	100.00	70.00	65.00	70.00	75.00
2	122271	1.51	2.18	2.36	24.34	25.00	25.66	34.21	1.32	1.32	1.32	0.00	23.68	22.37	26.97	21.05
3	1271	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
4	10337	100.00	100.00	100.00	82.35	88.24	100.00	94.12	100.00	100.00	100.00	100.00	88.24	94.12	82.35	94.12
5	629	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
6	3019	44.85	44.85	44.15	33.33	66.67	66.67	100.00	33.33	33.33	33.33	100.00	33.33	100.00	33.33	33.33
7	2709	100.00	100.00	68.66	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	50.00	50.00	50.00	100.00
8	16026	42.64	42.64	44.71	65.00	55.00	55.00	90.00	45.00	45.00	45.00	100.00	65.00	35.00	40.00	45.00
9	457	100.00	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	4830	65.94	65.94	63.66	62.50	62.50	50.00	100.00	62.50	62.50	62.50	100.00	75.00	75.00	75.00	62.50

Table C

Variable: Country of birth - ethnic

		Percentage of Records Incorrectly Imputed										
		Misclass. Rate		Frequency Distribution			Highest Probability			Nearest Neighbour		
Cat.	Records	10 TN	18 TN	10 TN	18 TN	N TRE	10 TN	18 TN	N TRE	10 TN	18 TN	N TRE
1	112643	0.11	0.12	22.56	18.80	24.81	0.00	0.00	0.00	24.06	11.28	21.80
2	19484	100.00	100.00	88.24	64.71	100.00	100.00	100.00	100.00	82.35	88.24	88.24
3	2091	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	3740	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
5	10693	100.00	100.00	100.00	85.71	100.00	100.00	100.00	100.00	100.00	71.43	85.71
6	257	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	22	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	636	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	431	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	109	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	2634	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	474	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	1538	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	14863	98.79	98.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	1339	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16	995	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	1428	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	703	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19	200	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20	2956	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

It can be noticed from these tables that there seems to be a relationship between the misclassification rates obtained from the tree and the percentage of records incorrectly imputed by categories. Both the misclassification rate and the percentage of records incorrectly imputed by categories tend to follow similar patterns most of the time. However, this relationship does not seem to be as strong as in the univariate case. It can also be observed that this relationship will not apply for the case where trees are not used.

This could be an important finding from the point of view of accuracy. It could be predicted from the tree by using the misclassification rate by categories, which categories of the variable being imputed will be more accurate than others after the imputation is done.

Another interesting finding obtained from this table is that in the case when trees are not used, the percentage of records incorrectly imputed by categories is usually higher (or at least equal) than the percentage of records incorrectly imputed when trees are used for the Frequency Distribution and Highest Probability methods.

Moreover, depending on the imputation method used, the percentage of records incorrectly imputed obtained from the case where trees are not used can be near to 100% for most of the categories as in the case of Highest Probability method where only one category is used for imputation. This corroborates the statement made previously that the use of trees improves the performance of the imputation depending on the method used.

In the case of Nearest Neighbour method, all the information, percentage of records incorrectly imputed using trees, percentage of records incorrectly imputed when trees are not used and misclassification rate, have more similar results than the rest of the methods when comparing different variables.

Again, this implies that there is not an impact on the imputation results when Nearest Neighbour method together with classification trees is used for imputation.

A set of graphs obtained from *Tables 6.8.4.1* can be found in the *Appendix 9.*

These graphs present the percentage of records incorrectly imputed using and not using trees and the misclassification rate by categories for the different imputation methods and different tree sizes. These graphs show how the lines for the percentage of records incorrectly imputed obtained using trees (red line) and the misclassification rates for the same categories (blue line) follow similar patterns. Alternatively, the line representing the percentage of records incorrectly imputed in the case where trees are not used (yellow line) is different from the two lines mentioned before.

6.8.5 Assessment of Biases and Variances

Since the variables used in the multivariate case are composite variables (which are treated as a simple categorical variable) and since the imputation for these simulations was carried out only for the case where the variables are missing at the same time and not individually, the estimator for the variance in the case of Frequency Distribution and Nearest Neighbour imputation methods presented in *Section 3.7* in *Chapter 3* is still valid. In the case where the missing information is present not only for the combination of variable but also for the individual variables forming that combinations (combination 1 and 2 in *Table 5.6.3* in *Chapter 5*), the estimator of the variance will need to be reformulated. This case is not assessed in this thesis.

Additionally, due to time constraints, the estimator of the variance in this chapter is only obtained for the Frequency Distribution method and combination of two variables missing at the same time only.

This section presents a set of summary tables; more detailed information can be seen in *Appendix 3*. It is important to point out that in here, as well as in the univariate case, the simulations were carried out using only one tree size for each combination of variables, since we had said before (in *Section 6.8.1*) that it seems to be no considerable differences on the results when using different tree sizes.

Tables 6.8.5.1 contain the information related to the simulations carried out for the biases, variances and variance estimation in the multivariate case for the case of Frequency Distribution imputation method.

Tables 6.8.5.1

Biases, Variances, Expected Variance Estimators and Coverage for the Frequency Distribution Imputation Method for the multivariate case

Table A

Composite variable: Ethnic - Country of birth

	Categories									
	1	2	3	4	5	6	7	8	9	10
$E(\hat{Y})$	126363	21847	2363	4185	12009	297	26	712	489	123
$E(\hat{Y})-Y$	0.00	-1.00	0.00	-1.00	-1.00	0.00	0.00	-1.00	0.00	0.00
$((E(\hat{Y})-Y)/Y)*100$	0.000	-0.004	0.000	-0.023	-0.008	0.000	0.000	-0.140	0.000	0.000
$E(\hat{V})$	35.00	14.00	2.00	3.00	9.00	0.00	0.00	1.00	0.00	0.00
$V(\hat{Y})$	36.00	15.00	2.00	4.00	10.00	0.00	0.00	1.00	0.00	0.00
$MSE(\hat{Y})$	36.00	16.00	2.00	5.00	11.00	0.00	0.00	2.00	0.00	0.00
Coverage	94.80	94.10	94.30	91.80	90.90	79.20	98.20	88.10	95.50	91.40

	11	12	13	14	15	16	17	18	19	20
$E(\hat{Y})$	2922	525	1730	16673	1509	1101	1597	781	226	3277
$E(\hat{Y})-Y$	-1.00	0.00	0.00	-1.00	0.00	-1.00	0.00	-1.00	0.00	7.00
$((E(\hat{Y})-Y)/Y)*100$	-0.034	0.000	0.000	-0.006	0.000	-0.090	0.000	-0.127	0.000	0.214
$E(\hat{V})$	2.00	0.00	1.00	12.00	1.00	1.00	1.00	1.00	0.00	3.00
$V(\hat{Y})$	3.00	0.00	1.00	12.00	1.00	1.00	1.00	1.00	0.00	7.00
$MSE(\hat{Y})$	4.00	0.00	1.00	13.00	1.00	2.00	1.00	2.00	0.00	56.00
Coverage	88.90	95.30	95.00	94.10	96.30	79.30	95.70	88.40	82.80	6.00

Table B

Composite variable: Ethnic - Long Term Illness

	Categories							
	1	2	3	4	5	6	7	8
$E(\hat{Y})$	17920	124268	3743	35979	739	6307	811	8989
$E(\hat{Y})-Y$	-1.00	0.00	-1.00	-1.00	-1.00	0.00	-1.00	5.00
$((E(\hat{Y})-Y)/Y)*100$	-0.005	0.000	-0.026	-0.002	-0.135	0.000	-0.123	0.055
$E(\hat{V})$	29.00	78.00	7.00	51.00	2.00	11.00	2.00	17.00
$V(\hat{Y})$	31.00	79.00	8.00	49.00	2.00	11.00	2.00	21.00
$MSE(\hat{Y})$	32.00	79.00	9.00	50.00	3.00	11.00	3.00	46.00
Coverage	94.00	95.00	93.90	95.30	91.80	93.80	86.00	77.00

Table C

Composite variable: Country of birth - Long term illness

	Categories									
	1	2	3	4	5	6	7	8	9	10
$E(\hat{Y})$	17553	137207	1452	11591	694	3365	3014	17999	500	5380
$E(\hat{Y})-Y$	0.00	0.00	-2.00	-1.00	0.00	-1.00	-1.00	-1.00	-1.00	6.00
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	0.000	0.000	-0.137	-0.008	0.000	-0.029	-0.033	-0.005	-0.199	0.111
$E(\hat{V})$	14.00	34.00	2.00	12.00	1.00	3.00	3.00	11.00	1.00	5.00
$V(\hat{Y})$	14.00	35.00	2.00	13.00	1.00	3.00	3.00	12.00	1.00	7.00
$MSE(\hat{Y})$	14.00	35.00	6.00	14.00	1.00	4.00	4.00	13.00	2.00	43.00
Coverage	95.00	94.70	79.00	93.70	90.80	91.40	91.50	92.60	88.60	24.70

It can be seen from these tables that the estimator of the total is approximately unbiased for all the categories of the target variables. That is, the difference between the real values of the total of records in this case and the expected values of the estimator of that total obtained by the simulations are very close to zero in the case of Frequency Distribution method. These results correspond to the findings obtained in the theoretical formulation.

Only few cases, category 20 for the composite variable Ethnic-Country of Birth, category 8 for composite variable Ethnic-Long term Illness and category 10 for composite variable Country of Birth-Long Term Illness, show slightly higher bias than the rest. However, these biases do not seem to be of major importance compared to the size of the values of the point estimates.

In terms of variability, we can see some differences in the values of the variance depending on the different categories of the target variables. Smaller variances are always for the smaller categories and bigger variances are always for the larger categories. However, even for the largest values we can see that the sizes of those variances are very small compared to the sizes of the point estimates.

In terms of estimation of the variance, we can see that this estimator basically lead to unbiased estimates. That is, there are not major differences between the expected value of the estimator of the variance and the actual variance.

The coverage, as in the univariate case, was estimated using a 95% nominal coverage level. This seems to be very stable in general, except for some cases with lower coverage, e.g. 77.00 the lowest.

Also, there are only few cases for which the coverage does not seem to be at the same level as the rest. In these cases, category 20 for the composite variable Ethnic-Country of Birth and category 10 for composite variable Country of Birth-Long Term Illness, we can see that the bias is bigger than the variance, which makes the coverage lower. That is, the fact the confidence intervals are moved to one side (right side in this cases) given the values of the bias and the variances are small, make the coverage poor.

However, if we take the value of the Mean Square Error for these categories and divide them by their respective totals in each case, we can see that the estimator is very accurate since the value of the relative Mean Square Errors are very smalls.

Tables 6.8.5.2 contain the information for the bias, variance and coverage for the case of Highest Probability imputation method. In the case of the coverage, this was estimated using a 95% nominal coverage level and the variance obtained from the simulations given that estimates for these variances were not obtained.

Tables 6.8.5.2

Biases, Variances and Coverage for the Highest Probability Imputation Method for the multivariate case

Table A

Composite variable: Ethnic - Country of birth

	Categories									
	1	2	3	4	5	6	7	8	9	10
$E(\hat{Y})$	126420	21830	2361	4183	12001	297	26	712	489	123
$E(\hat{Y})-Y$	57.00	-18.00	-2.00	-3.00	-9.00	0.00	0.00	-1.00	0.00	0.00
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	0.045	-0.082	-0.084	-0.071	-0.074	0.000	0.000	-0.140	0.000	0.000
$V(\hat{Y})$	37.00	16.00	2.00	3.00	10.00	0.00	0.00	1.00	0.00	0.00
$MSE(\hat{Y})$	3286.00	340.00	6.00	12.00	91.00	0.00	0.00	2.00	0.00	0.00
Coverage	0.00	0.00	71.3	58.3	18.7	79.3	98.2	88.1	68.7	91.4

	11	12	13	14	15	16	17	18	19	20
$E(\hat{Y})$	2921	525	1729	16661	1508	1101	1596	781	226	3267
$E(\hat{Y})-Y$	-2.00	0.00	-1.00	-13.00	-1.00	-1.00	-1.00	-1.00	0.00	-3.00
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	-0.068	0.000	-0.057	-0.077	-0.066	-0.090	-0.062	-0.127	0.000	-0.091
$V(\hat{Y})$	2.00	0.00	1.00	12.00	1.00	1.00	1.00	1.00	0.00	3.00
$MSE(\hat{Y})$	6.00	0.00	2.00	181.00	2.00	2.00	2.00	2.00	0.00	12.00
Coverage	59.1	65.8	59.7	2.7	64.3	78.7	66.4	88.4	82.8	74.8

Table B

Composite variable: ETHNIC -LTILL

	Categories							
	1	2	3	4	5	6	7	8
$E(\hat{Y})$	17885	124364	3738	35941	738	6302	810	8977
$E(\hat{Y})-Y$	-36.00	96.00	-6.00	-39.00	-2.00	-5.00	-2.00	-7.00
$((E(\hat{Y})-Y)/Y)*100$	-0.200	0.077	-0.160	-0.108	-0.270	-0.079	-0.246	-0.077
$V(\hat{Y})$	41.00	100.00	9.00	65.00	2.00	12.00	2.00	21.00
$MSE(\hat{Y})$	1337.00	9316.00	45.00	1586.00	6.00	37.00	6.00	70.00
Coverage	0	0	51.9	0.3	77.1	63.9	71.7	61.4

Table C

Composite variable: Country of birth - Long term illness

	Categories									
	1	2	3	4	5	6	7	8	9	10
$E(\hat{Y})$	17537	137247	1452	11579	693	3366	3012	17997	500	5373
$E(\hat{Y})-Y$	-16.00	40.00	-2.00	-13.00	-1.00	0.00	-3.00	-3.00	-1.00	-1.00
$((E(\hat{Y})-Y)/Y)*100$	-0.091	0.029	-0.137	-0.112	-0.144	0.000	-0.099	-0.016	-0.199	-0.018
$V(\hat{Y})$	18.00	41.00	2.00	13.00	1.00	4.00	3.00	13.00	1.00	6.00
$MSE(\hat{Y})$	274.00	1641.00	6.00	182.00	2.00	4.00	12.00	22.00	2.00	7.00
Coverage	2.5	0	77.6	5.8	82.5	91.8	54.7	88.2	88.2	90.3

It can be noticed from these tables that the Highest Probability imputation method is not an unbiased procedure as demonstrated in theory in *Chapter 5*. However, the biased introduced by the method does not seem to be very high with respect to the values of the totals. Moreover, some of the bias are close to or even zero.

It can also be seen that the bias are always positive for the major category and negative for the others as would be expected since the method uses the value with highest frequency to impute the missing records.

In terms of variability we can see that the values of the variances are larger than the values of the variances obtained in the Frequency Distribution case. Additionally, it can be noticed that the bigger variances are always for the larger categories. Estimates for those variances were not obtained in this case.

The coverage values obtained in this case do not seem to be as good as in the Frequency Distribution case. It can be noticed from the tables that some of the variable have zero

coverage while some others have better results, for example 98% coverage. There seems to be a much larger variation on the coverage values than in the Frequency Distribution case.
As explained in the univariate case, a reason for this coverage problem in the Highest Probability case is that this method is not an unbiased procedure. We can see that the size of the bias is as big as the size of the variance, which does not occur in the Frequency Distribution case.
Additionally, it can be seen that the smaller coverage is always present in the categories with bigger number of records. As explained before, that occurs because of the size of the bias in those categories.

As in the univariate case, there are some general findings we summarise hereafter in terms of the analysis

The use of classification trees does improve the performance of the imputation. This improvement can be noticed from the point of view of the maintenance of marginal distributions (either individual marginal or joint distributions) in most of the cases but from the point of view of percentage of records correctly imputed.

As in the univariate case, the use of different tree sizes does not have a major impact on those results. Moreover, the use of the optimal tree given by CART does not make much difference on the results.

Frequency Distribution and Nearest Neighbour methods preserve marginal distributions while Highest Probability does not. However, Highest Probability is the best performing imputation method in terms of preservation of individual values.
In the case of Nearest Neighbour, the results are very similar in both cases (when using and not trees). That is, in general, the use of trees does not seem to have a major impact on the results when using Nearest Neighbour procedure.
As a general conclusion we can say that Frequency Distribution is the best performing method overall as it preserves marginal distributions, has a reasonable level of preservation of individual values, produces unbiased estimates for the total and has the lowest variability between all the methods.

Frequency Distribution and Nearest Neighbour methods produce unbiased estimates for the total number of records in a specific category. In contrast, Highest Probability method does not lead to unbiased estimates as shown in the theoretical results.
In terms of variability, Frequency Distribution and Highest Probability methods produce very similar variances. However, the case of Nearest Neighbour does not present the same results. It can be seen that this last method produces larger variances than the rest of the methods.

If the results for the mean square error obtained from the different methods are compared, we can see that the lowest values are always found for Frequency Distribution followed by Nearest Neighbour and Highest Probability (due to the bias) respectively.

As the estimation of the variance in the multivariate case was only carried out for the Frequency Distribution methods, no comparison can be made. However, the simulations presented for this case show that this method basically leads to unbiased estimates for the variance confirming the theoretical results.

It has been shown in the results that the coverage, is more variable that in the univariate case. However, the values are very high in most of the cases.

CHAPTER 7

SUMMARY AND CONCLUSIONS

7.1. SUMMARY

Censuses are the most important statistical demographic operation carried out by any country. As any statistical collection processes, censuses are susceptible to "nonresponse". Nonresponse occurs when any investigated variable for any element within the "universe of study" is missing in the final format for the analysis. Nonresponse can affect analysis, leading to erroneous or invalid findings and consequent decision-making.

The deficiencies of the current methods actually used for solving the problem of missing information in census (described in *Chapter 1*), added to the importance of the census data for the statistics in a country, are the main reasons why research about improved methodology for imputing this kind of data has been undertaken.
The main idea was to investigate an alternative method, which uses a different approach to the current methods available, being also simple and efficient.

The method investigated in this research involved the use of classification trees as a first step, followed by imputation using common methods for categorical data within each imputation class (terminal nodes of the tree).

The classification technique used in this research is called "Classification and Regression Trees" (CART). CART technique is basically a set of classification rules (recursive binary segmentation) that partition the data set into mutually exhaustive and non-overlapping subsets (terminal nodes) based on the values of a group of explanatory variables. These subsets are expected to be internally more homogeneous with respect to response variable (variable for which the tree is generated) than the whole database.
Once the classification is made, each imputation method is applied independently within each terminal node. Three different basic imputation methods for categorical data are

implemented in this thesis in order to compare their results given the classification. The selection of the methods included Probability Distribution imputation, Highest Probability (Modal) imputation as well as the use of Nearest Neighbour procedure as it seems to be a common factor in most of the new methodologies created for census data, as mentioned in *Chapter 1*.

The combination of classification and imputation allow for the measure of: 1) the effect of using this classification technique on the imputation results (including the use of different tree-sizes), and 2) the accuracy of the different imputation methods based on this classification technique.

The analysis was carried out for two different targets: the univariate case where a single variable is imputed, and the multivariate case where a composite variable is imputed. A composite variable is defined by the cross-classification of two or more single variables. The use of the composite variable allows for the imputation of two or more single variables at the same time.

Preservation of joint and individual marginal distributions as well as preservation of individual values are evaluated (comparing imputed values against real values). Graphs and tests for those comparisons are presented. Additionally, assessment of biases and variances, as well as variance estimation in some cases, are also presented.

The simulation was made using a subset of UK 1991 Census information. Only categorical variables related to persons (except age, which was converted to categorical) were used for the analysis. After deleting the records with missing information from the original database, artificial holes were created using the real pattern of missing information present in the original database. This allowed for the measure of the accuracy of the imputation by comparing the real and the imputed values.

7.2. GENERAL CONCLUSIONS

Some important conclusion can be obtained from the research presented in this thesis. We divide the conclusion in different sections, as follows:

7.2.1 Using Classification Trees

We conclude that, in general, the use of classification trees does improve the performance of the imputation. As seen on the results this improvement cannot be seen from the point of view of the maintenance of marginal distributions (either individual marginal or joint distributions) in most of the cases but from the point of view of percentage of records correctly imputed.

A comparison between the percentage of records correctly imputed show that there are considerable differences from the case where trees are not used, specially in for the Frequency Distribution method.

In the case of Highest Probability method, even when using trees does not allow for the maintenance of the marginal distributions, it does improve the actual distribution as it imputes values depending on the modal category of the terminal nodes while in the case where trees are not used, the imputation is carried out using the modal category of the whole database. That is, using trees produce a closer shape to the real distribution than no using trees. Therefore, we can say that using tree does improve the recreation of marginal distributions even when the values of the Wald statistic show that they are not preserved. Additionally, the use of the tree also increases most of the times the percentage of the number of records correctly imputed when using this imputation method.

7.2.2. Comparing Different Tree Sizes

It has been demonstrated that, even when there are differences in the results when using or not trees (i.e. trees improve the performance of the imputation results), the use of different tree sizes does not have a major impact on those results. The simulations carried out show a stable behaviour across the different tree sizes, even in terms of misclassification rates obtained from the tree process.

Moreover, the optimal tree given by CART is meant to be optimal in terms of complexity (number of terminal nodes) and misclassification rate. In this sense, the use of the optimal tree could be expected to give the best performance, however, it can be observed from the results that this hypothesis is not necessarily correct. Using the optimal tree given by CART does not make much difference on the results.

7.2.3. Comparing Imputation Methods

In terms of comparisons between imputation methods, we have shown that both Frequency Distribution and Nearest Neighbour methods preserve marginal distributions while Highest Probability does not. However, the results from the simulations show that in terms of preservation of individual values, Highest Probability is the best performing imputation method with a minimum of almost 60% of individual values preserved in the worst case and a maximum of almost 90%.

The simulations carried out also show that in the case of Nearest Neighbour, the results are very similar in both cases (when using and not trees). That is, in general, the use of trees does not seem to have a major impact on the results when using Nearest Neighbour procedure. One of the reasons why this happened could be the fact that by definition Nearest Neighbour procedure look for the closest donor, which should be found with or without the use of classification.

As a general conclusion we can say that Frequency Distribution is the best performing method overall as it preserves marginal distributions, has a reasonable level of preservation of individual values, produces unbiased estimates for the total and has the lowest variability between all the methods.

7.2.4. Bias and Variance

It has been shown that both Frequency Distribution and Nearest Neighbour methods produce unbiased estimates for the total number of records in a specific category. In contrast, Highest Probability method does not lead to unbiased estimates as shown in the theoretical results. This bias in the case of Highest Probability method is always positive for the major categories and negative for the minor ones, overestimating the categories with more units and underestimating the categories with less number of units.

In terms of variability, we can see that the values for the variances in the case of Frequency Distribution and Highest Probability methods are very similar. However, the case of Nearest Neighbour does not present the same results. It can be seen that this last method produces larger variances than the rest of the methods.

If the results for the mean square error obtained from the different methods are compared, we can see that the lowest values are always found for Frequency Distribution followed by Nearest Neighbour and Highest Probability (due to the bias) respectively.

7.2.5. Variance Estimation

The simulations for the estimation of the variance, as well as the theoretical formulation, were carried out for Frequency Distribution and Nearest Neighbour methods and not for Highest Probability method.

The results of the estimation confirm the theoretical result that the estimator of the variance proposed for the Frequency Distribution case is an unbiased estimator. However, even when in theory the estimator of the variance for the Nearest Neighbour seems to be unbiased, some differences between the real variance (value assumed as real in the simulations) and the expected values over the 1000 samples of the estimator of the variance can be found in few cases. It has to be said that these differences are probably big when estimating the variance but they are not very important in terms of the values of the point estimates as they are very small with respect to these values.

7.2.6. Coverage

It has been shown in the results that the coverage, i.e. the proportion of confidence intervals that contain the parameter, is over 94% all the time in the univariate case for both Frequency Distribution and Nearest Neighbour and very high most of the times in the multivariate case for the Frequency Distribution case.

7.2.7. Comparison with Hot Deck Imputation

The simulations for the Sequential Hot Deck were carried out only for the univariate case. Comparisons between the proposed method and the method proposed in this thesis show that in terms of the point estimates any of the Frequency Distribution, Nearest Neighbour and Hot Deck produces unbiased estimators. In terms of variability, the sequential Hot Deck method produces larger variances than any of the imputation procedures investigated in this research.

However, if a comparison between the mean square errors is made, we can see that sequential Hot Deck performs better than the Highest Probability procedure, producing smaller MSE.

Thus, Frequency Distribution is still the best performing imputation methods in this research, followed by Nearest Neighbour, Sequential Hot Deck and Highest Probability respectively.

7.3. PROS AND CONS OF THE PROPOSED METHODOLOGY

There are some advantages of the proposed methodology:

- The proposed method involves the use of classification as a first step. The aim of classification is to ensure the selection of the best possible donor, since both the donor and the recipient come from the same imputation class, which guarantees same characteristics (same values) for the observed variables for both of the records.

- One of the main aspects concerning the proposed approach is the maintenance of joint distributions, which means upholding correlations between variables when working on the multivariate missing case. The method proposed allows for this aspect since the imputation will be made jointly for all the missing variables belonging to a specific record trying to obtain those imputations from the same donor.

- Another advantage of the proposed approach is the fact that it does not imply the use of complicated procedures or sophisticated technical resources. This new method is easy to implement and does not require a large amount of time.

- The use of as many variables as possible (as many as are involved in the relationship) in the classification step is another advantage of this proposal. They guarantee upholding relationships between variables as well as defining very well the groups from which the donors are going to be taken. This also makes the selection of the donor easier and faster, since this is sought in that specific class and not in any other.

- The method proposed allows for the use of missing covariates in the classification process, which is not normally the case when using other procedures. That is, the records containing missing information for the auxiliary variables can be included in the process of growing the tree. The inclusion of those records permits the use of as much information as available, which could be crucial at times when the information present is not sufficient.

  The classification, of records with missing values for the auxiliary variables, is made by using alternative classifiers called surrogates, which is basically another x_{ik} variable (which value is present in that record) correlated with the one missing one that classifies the records in the same way (or very similar way) as the original classifier.

- Besides the use of surrogates as classifiers, another potential advantage of their use is the possibility of imputing several missing values present in a single record in a sequential way. This represents a significant issue about using classification techniques together with imputation procedures. However, this aspect was not study in this work.

There are also some disadvantages in the proposed methodology that can be mentioned.

- As explained in *Chapter* 2, *4* and *6*, the classification tree is created for a single variable (or composite variable in the multivariate case). That implies, in a very strict sense, that each combination of missing information requires its own classification tree for the imputation process. This is a very difficult task to achieve since, as shown in *Appendix 1*, only with 8 variables we have 168 combinations of missing values. Therefore, in order to reduce time consuming and complexity, imputations are carried out with just few classification trees (sometimes even just one). That is, a classification tree created for a specific missing combination has to be also used for other variables.
  This can be seen as a disadvantage since the classification is not specially created for the variable to be imputed, and somehow, it could be not the best classification for that specific variable. However, when variables are highly correlated, a classification tree created for a composite variable can be perfectly used for the single variables involved in that combination. A solution for choosing a combination for generating the tree to be used could be the use the largest combination in term of percentage of missing information.

- Another disadvantage of the proposal is the fact that the results are not generated instantaneously (directly) from a software. That is, CART (software) creates the classification tree. After the tree is generated, the set of rules defining the terminal nodes are used to create a computing program for dividing the population and carrying out with the imputation process using a different software (FoxPro in this case). Therefore, programming can be one of the biggest disadvantages of the proposed approach.

- As well as for the classification and imputation process, the bias and variance assessment have to be programmed, including the estimation of the variance, if wanted.

7.4. FURTHER RESEARCH

This thesis represents just the first stage on the research of the use of classification trees for missing item imputation. Further research should be done in order to assess more aspects about this subject. The work can be divided into two different parts, univariate case and multivariate case.

In terms of the univariate case, there are basically three points that can be done. These are:

- The estimation of the variance in the case of the Highest Probability imputation methods, although this is not of great importance since it was shown that this method does not lead to unbiased estimation.

- A comparison between the proposed methods and other new imputation methodology such as DIS (from the Office for National Statistics in the UK).

- A comparison using alternative imputation methods like Logistic or Log-linear regression.

In the multivariate case, the research presented in this thesis does not cover some important aspects related to the subject. There are certain points that should be studied further. These include:

- Formulation of an estimator for the variance in the Frequency Distribution case when using classification trees. In the case of Nearest Neighbour, the estimation of the variance does not seem very important since the use of the classification trees does not have a major impact on the results.

- Even when this thesis defines a way in which surrogates can be used for imputation, no further research (neither theoretical nor empirical) was made about this matter. A separate study is recommended in order to assess the viability and properties of this procedure.

- Since this research does not investigate any existing method in the multivariate case, comparisons between the proposed method and different methods have to be done in order to evaluate the relative merits of the proposed method. The most reasonable comparisons would be the use of Hot Deck imputation in multivariate missing data and new methodologies such as DIS (from the Office for National Statistics in the UK) as in the univariate case.

APPENDIX

Appendix 1

Missing Information Pattern

age	alwprim	cob	ethnic	ltill	marcon	sex	welsh	Total	perct
							■	3230	13.39
						■		1190	4.93
						■	■	38	0.16
					■			1701	7.05
					■		■	116	0.48
					■	■		32	0.13
					■	■	■	2	0.01
				■				3224	13.37
				■			■	347	1.44
				■		■		39	0.16
				■		■	■	4	0.02
				■	■			106	0.44
				■	■		■	13	0.05
				■	■	■		1	0.00
				■	■	■	■	1	0.00
			■					3916	16.24
			■				■	156	0.65
			■			■		45	0.19
			■			■	■	2	0.01
			■		■			76	0.32
			■		■		■	11	0.05
			■		■	■		2	0.01
			■	■				521	2.16
			■	■			■	88	0.36
			■	■		■		12	0.05
			■	■		■	■	1	0.00
			■	■	■			22	0.09
			■	■	■		■	1	0.00
			■	■	■	■		3	0.01
		■						1751	7.26
		■					■	128	0.53
		■				■		23	0.10
		■				■	■	1	0.00
		■			■			52	0.22
		■			■		■	9	0.04
		■			■	■		1	0.00
		■		■				252	1.04
		■		■			■	71	0.29
		■		■		■		1	0.00
		■		■		■	■	1	0.00
		■		■	■			11	0.05
		■		■	■		■	2	0.01
		■	■					178	0.74
		■	■				■	23	0.10
		■	■			■		7	0.03
		■	■			■	■	2	0.01
		■	■		■			7	0.03
		■	■		■		■	1	0.00
		■	■	■				520	2.16
		■	■	■			■	190	0.79
		■	■	■		■		22	0.09
		■	■	■		■	■	3	0.01
		■	■	■	■			24	0.10
		■	■	■	■		■	11	0.05
		■	■	■	■	■		1	0.00
		■	■	■	■	■	■	1	0.00
	■							1406	5.83
	■						■	101	0.42
	■					■		15	0.06
	■					■	■	2	0.01
	■				■			78	0.32
	■				■		■	16	0.07
	■				■	■		3	0.01
	■				■	■	■	1	0.00
	■			■				203	0.84
	■			■			■	72	0.30
	■			■		■		2	0.01
	■			■	■			19	0.08
	■			■	■		■	8	0.03
	■			■	■	■		1	0.00
	■		■					75	0.31
	■		■				■	18	0.07
	■		■			■		3	0.01
	■		■			■	■	1	0.00
	■		■		■			10	0.04
	■		■		■		■	3	0.01
	■		■	■				64	0.27
	■		■	■			■	35	0.15
	■		■	■	■			17	0.07
	■		■	■	■		■	4	0.02
	■	■						77	0.32
	■	■					■	12	0.05
	■	■				■	■	1	0.00
	■	■			■			7	0.03

age	alwprim	cob	ethnic	ltill	marcon	sex	welsh	Total	perct
	■	■			■		■	2	0.01
	■	■		■				33	0.14
	■	■		■			■	21	0.09
	■	■		■		■		2	0.01
	■	■		■		■	■	1	0.00
	■	■		■	■			10	0.04
	■	■		■	■		■	4	0.02
	■	■		■	■	■		1	0.00
	■	■	■					23	0.10
	■	■	■				■	2	0.01
	■	■	■		■			2	0.01
	■	■	■		■		■	5	0.02
	■	■	■	■				399	1.65
	■	■	■	■			■	100	0.41
	■	■	■	■		■		7	0.03
	■	■	■	■		■	■	1	0.00
	■	■	■	■	■			77	0.32
	■	■	■	■	■		■	78	0.32
	■	■	■	■	■	■	■	2	0.01
■								1737	7.20
■							■	47	0.19
■						■		83	0.34
■						■	■	10	0.04
■					■			91	0.38
■					■		■	11	0.05
■					■	■		18	0.07
■					■	■	■	3	0.01
■				■				63	0.26
■				■			■	7	0.03
■				■		■		9	0.04
■				■		■	■	4	0.02
■				■	■			14	0.06
■				■	■		■	3	0.01
■			■					68	0.28
■			■				■	3	0.01
■			■			■		1	0.00
■			■			■	■	2	0.01
■			■		■			6	0.02
■			■		■		■	1	0.00
■			■		■	■		3	0.01
■			■	■				20	0.08
■			■	■			■	6	0.02
■			■	■		■		1	0.00
■			■	■		■	■	2	0.01
■			■	■	■			3	0.01
■			■	■	■	■		1	0.00
■		■						67	0.28
■		■					■	8	0.03
■		■				■		5	0.02
■		■			■			8	0.03
■		■			■		■	1	0.00
■		■		■				29	0.12
■		■		■			■	10	0.04
■		■		■		■	■	8	0.03
■		■		■	■			2	0.01
■		■		■	■		■	3	0.01
■		■	■					13	0.05
■		■	■				■	1	0.00
■		■	■		■			2	0.01
■		■	■	■				116	0.48
■		■	■	■			■	77	0.32
■		■	■	■		■		30	0.12
■		■	■	■		■	■	9	0.04
■		■	■	■	■			3	0.01
■		■	■	■	■		■	3	0.01
■		■	■	■	■	■		2	0.01
■		■	■	■	■	■	■	176	0.73
■	■							55	0.23
■	■						■	1	0.00
■	■					■		1	0.00
■	■				■			4	0.02
■	■				■	■		1	0.00
■	■			■				8	0.03
■	■			■		■	■	1	0.00
■	■		■					2	0.01
■	■		■	■				2	0.01
■	■	■						6	0.02
■	■	■					■	2	0.01
■	■	■			■			1	0.00
■	■	■		■				1	0.00
■	■	■	■		■			1	0.00
■	■	■	■	■				12	0.05
■	■	■	■	■			■	8	0.03
■	■	■	■	■		■		3	0.01
								24116	100.00

Appendix 2

Bias and Variance of the estimator of the total depending on the number of terminal used for the variable Ethnic by categories

2 Terminal Nodes

	Categories			
	1	2	3	4
Y	142189	39724	7047	9796
$E(\hat{Y})$	142189	39724	7047	9796
$E(\hat{Y})-Y$	0.00	0.00	0.00	0.00
$S(\hat{Y})$	22	20	10	12
$V(\hat{Y})$	465	403	97	151

3 Terminal Nodes

	Categories			
	1	2	3	4
Y	142189	39724	7047	9796
$E(\hat{Y})$	142189	39725	7047	9796
$E(\hat{Y})-Y$	0.00	1.00	0.00	0.00
$S(\hat{Y})$	22	19	10	12
$V(\hat{Y})$	463	365	96	141

4 Terminal Nodes

	Categories			
	1	2	3	4
Y	142189	39724	7047	9796
$E(\hat{Y})$	142189	39724	7047	9796
$E(\hat{Y})-Y$	0.00	0.00	0.00	0.00
$S(\hat{Y})$	22	19	9	12
$V(\hat{Y})$	465	366	80	133

10 Terminal Nodes

	Categories			
	1	2	3	4
Y	142189	39724	7047	9796
$E(\hat{Y})$	142189	39724	7046	9797
$E(\hat{Y})-Y$	0.00	0.00	-1.00	1.00
$S(\hat{Y})$	22	19	9	12
$V(\hat{Y})$	464	366	80	136

13 Terminal Nodes

	Categories			
	1	2	3	4
Y	142189	39724	7047	9796
$E(\hat{Y})$	142189	3974	7046	9797
$E(\hat{Y})-Y$	0.00	0.00	-1.00	1.00
$S(\hat{Y})$	22	19	9	12
$V(\hat{Y})$	464	366	80	136

Appendix 3

Estimator Properties and Performance Indicators

Frequency Distribution Method

	ETHNIC				COUNTRY OF BIRTH					LONG-TERM ILLNESS	
	Categories				Categories					Categories	
	1	2	3	4	1	2	3	4	5	1	2
Y	142189	39724	7047	9796	154760	13046	4060	21015	5875	23217	175539
$E(\hat{Y})$	142189	39724	7046	9797	154760	13046	4060	21015	5876	23217	175539
$E(\hat{Y})-Y$	-0.34	0.03	-1.07	1.37	0.36	-0.36	-0.40	-0.47	0.87	0.16	-0.16
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	0.000	0.000	-0.015	0.013	0.000	-0.002	-0.009	-0.002	0.014	0.000	0.000
$E(\hat{S})$	22.13	20.49	9.23	11.72	12.96	9.73	4.51	8.31	6.09	15.25	15.25
$S(\hat{Y})$	21.54	19.13	8.93	11.67	12.68	9.71	4.39	8.02	6.17	15.06	15.06
$E(\hat{S})-S(\hat{Y})$	0.59	1.36	0.30	0.05	0.28	0.02	0.12	0.29	-0.08	0.19	0.19
$E(\hat{V})$	489.58	419.79	85.17	137.27	167.97	94.63	20.34	69.14	37.11	232.55	232.55
$V(\hat{Y})$	463.76	365.97	79.83	136.23	160.80	94.27	19.25	64.30	38.11	226.69	226.69
$E(\hat{V})-V(\hat{Y})$	25.82	53.82	5.34	1.04	7.17	0.36	1.09	4.84	-1.00	5.86	5.86
Coverage	94.90	95.60	94.80	95.00	95.80	95.50	95.10	96.60	95.10	95.40	95.40
$MSE(\hat{Y})$	463.88	365.97	80.97	138.11	160.93	94.40	19.41	64.52	38.87	226.72	226.72
$\sqrt{MSE(\hat{Y})}$	21.54	19.13	9.00	11.75	12.69	9.72	4.41	8.03	6.23	15.06	15.06
$\left(\sqrt{MSE(\hat{Y})}/Y\right)*100$	0.02	0.05	0.13	0.12	0.01	0.07	0.11	0.04	0.11	0.06	0.01
$\left(MSE(\hat{Y})/Y\right)*100$	0.33	0.92	1.15	1.41	0.10	0.72	0.48	0.31	0.66	0.98	0.13
$\left((E(\hat{V})-V(\hat{Y}))/V(\hat{Y})\right)*100$	5.567	14.706	6.689	0.763	4.458	0.381	5.662	7.527	-2.623	2.585	2.585
$\left((E(\hat{V})-V(\hat{Y}))/Y\right)*100$	0.018	0.135	0.075	0.010	0.004	0.002	0.026	0.023	-0.017	0.025	0.003

Appendix 3

Estimator Properties and Performance Indicators

Highest Probability Method

	ETHNIC				COUNTRY OF BIRTH					LONG-TERM ILLNESS	
	Categories				Categories					Categories	
	1	2	3	4	1	2	3	4	5	1	2
Y	142189	39724	7047	9796	154760	13046	4060	21015	5875	23217	175539
$E(\hat{Y})$	142653	39496	6993	9714	154915	12944	4058	20975	5864	22950	175806
$E(\hat{Y})-Y$	464.00	-328.00	-54.00	-82.00	155.00	-102.00	-2.00	-40.00	-11.00	-267.00	267.00
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	0.326	-0.825	-0.766	-0.837	0.100	-0.781	-0.049	-0.190	-0.187	-1.150	0.152
$S(\hat{Y})$	22.32	19.95	9.27	12.25	13.08	9.70	4.90	8.54	6.86	17.06	17.06
$V(\hat{Y})$	498.00	398.00	86.00	150.00	171.00	94.00	24.00	73.00	47.00	291.00	291.00
Coverage	0.00	0.00	0.00	0.00	0.00	0.00	92.20	0.10	66.20	0.00	0.00
$MSE(\hat{Y})$	215794.00	107982.00	3002.00	6874.00	24196.00	10498.00	28.00	1673.00	168.00	71580.00	71580.00
$\sqrt{MSE(\hat{Y})}$	464.54	328.61	54.79	82.91	155.55	102.46	5.29	40.90	12.96	267.54	267.54
$\left(\sqrt{MSE(\hat{Y})}/Y\right)*100$	0.33	0.83	0.78	0.85	0.10	0.79	0.13	0.19	0.22	1.15	0.15
$\left(MSE(\hat{Y})/Y\right)*100$	151.77	271.83	42.60	70.17	15.63	80.47	0.69	7.96	2.86	308.31	40.78

Appendix 3

Estimator Properties and Performance Indicators

Nearest Neighbour Method

	ETHNIC				COUNTRY OF BIRTH					LONG-TERM ILLNESS	
	Categories				Categories					Categories	
	1	2	3	4	1	2	3	4	5	1	2
Y	142189	39724	7047	9796	154760	13046	4060	21015	5875	23217	175539
$E(\hat{Y})$	142190	39724	7047	9795	154760	13046	4060	21015	5874	23217	175539
$E(\hat{Y})-Y$	0.64	-0.35	0.30	-0.58	0.21	0.36	-0.15	0.09	-0.50	0.10	-0.10
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	0.000	0.000	0.004	-0.005	0.000	0.002	-0.003	0.000	-0.008	0.000	0.000
$E(\hat{S})$	31.16	28.84	12.99	16.49	18.23	13.73	6.36	11.65	8.60	21.49	21.49
$S(\hat{Y})$	28.25	25.92	12.55	16.38	17.87	13.47	5.97	11.05	8.11	21.26	21.26
$E(\hat{S})-S(\hat{Y})$	2.91	2.92	0.44	0.11	0.36	0.26	0.39	0.60	0.49	0.23	0.23
$E(\hat{V})$	970.78	831.78	168.75	271.95	332.16	188.50	40.48	135.87	73.94	461.94	461.94
$V(\hat{Y})$	797.95	671.59	157.61	268.23	319.22	181.47	35.70	122.06	65.80	451.90	451.90
$E(\hat{V})-V(\hat{Y})$	172.83	160.19	11.14	3.72	12.94	7.03	4.78	13.81	8.14	10.04	10.04
Coverage	95.30	94.40	96.40	95.40	96.00	95.00	95.20	95.50	96.10	95.30	95.30
$MSE(\hat{Y})$	798.36	671.71	157.70	268.57	319.26	181.60	35.72	122.07	66.05	451.91	451.91
$\sqrt{MSE(\hat{Y})}$	28.26	25.92	12.56	16.39	17.87	13.48	5.98	11.05	8.13	21.26	21.26
$\left(\sqrt{MSE(\hat{Y})}/Y\right)*100$	0.02	0.07	0.18	0.17	0.01	0.10	0.15	0.05	0.14	0.09	0.01
$\left(MSE(\hat{Y})/Y\right)*100$	0.56	1.69	2.24	2.74	0.21	1.39	0.88	0.58	1.12	1.95	0.26
$\left((E(\hat{V})-V(\hat{Y}))/V(\hat{Y})\right)*100$	21.65	23.85	7.06	1.38	4.05	3.87	13.38	11.31	12.37	2.22	2.22
$\left((E(\hat{V})-V(\hat{Y}))/Y\right)*100$	0.121	0.403	0.158	0.037	0.008	0.053	0.117	0.065	0.138	0.043	0.005

Appendix 3

Estimator Properties and Performance Indicators

Hot Deck Method (Using 2 Variables for Classification)

	ETHNIC				COUNTRY OF BIRTH					LONG-TERM ILLNESS	
	Categories				Categories					Categories	
	1	2	3	4	1	2	3	4	5	1	2
Y	142189	39724	7047	9796	154760	13046	4060	21015	5875	23217	175539
$E(\hat{Y})$	142189	39724	7046	9797	154760	13046	4060	21015	5875	23217	175539
$E(\hat{Y})-Y$	0.00	0.00	-1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	0.000	0.000	-0.014	0.010	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$S(\hat{Y})$	31.00	27.00	15.00	17.00	21.00	13.00	7.00	16.00	9.00	22.00	22.00
$V(\hat{Y})$	961.00	729.00	225.00	289.00	441.00	169.00	49.00	256.00	81.00	484.00	484.00
$MSE(\hat{Y})$	961.00	729.00	226.00	290.00	441.00	169.00	49.00	256.00	81.00	484.00	484.00
$\sqrt{MSE(\hat{Y})}$	31.00	27.00	15.03	17.03	21.00	13.00	7.00	16.00	9.00	22.00	22.00
$\left(\sqrt{MSE(\hat{Y})}/Y\right)*100$	0.021	0.067	0.213	0.173	0.013	0.099	0.172	0.076	0.153	0.094	0.012
$\left(MSE(\hat{Y})/Y\right)*100$	0.675	1.835	3.207	2.960	0.284	1.295	1.206	1.218	1.378	2.084	0.275

Appendix 3

Estimator Properties and Performance Indicators

Hot Deck Method (Using 3 Variables for Classification)

	ETHNIC				COUNTRY OF BIRTH					LONG-TERM ILLNESS	
	Categories				Categories					Categories	
	1	2	3	4	1	2	3	4	5	1	2
Y	142189	39724	7047	9796	154760	13046	4060	21015	5875	23217	175539
$E(\hat{Y})$	142190	39724	7046	9796	154759	13046	4060	21015	5875	23217	175539
$E(\hat{Y})-Y$	1.00	0.00	-1.00	0.00	-1.00	0.00	0.00	0.00	0.00	0.00	0.00
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	0.000	0.000	-0.014	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$S(\hat{Y})$	32.00	28.00	14.00	17.00	21.00	13.00	7.00	16.00	9.00	20.00	20.00
$V(\hat{Y})$	1024.00	784.00	196.00	289.00	441.00	169.00	49.00	256.00	81.00	400.00	400.00
$MSE(\hat{Y})$	1025.00	784.00	197.00	289.00	442.00	169.00	49.00	256.00	81.00	400.00	400.00
$\sqrt{MSE(\hat{Y})}$	32.02	28.00	14.04	17.00	21.02	13.00	7.00	16.00	9.00	20.00	20.00
$\left(\sqrt{MSE(\hat{Y})}/Y\right)*100$	0.022	0.070	0.199	0.173	0.013	0.099	0.172	0.076	0.153	0.086	0.011
$\left(MSE(\hat{Y})/Y\right)*100$	0.720	1.973	2.795	2.950	0.285	1.295	1.206	1.218	1.378	1.722	0.227

Appendix 3

Estimator Properties and Performance Indicators

Frequency Distribution Method

	Country of Birth - Ethnic									
	Categories									
	1	2	3	4	5	6	7	8	9	10
Y	126363	21848	2363	4186	12010	297	26	713	489	123
$E(\hat{Y})$	126363	21847	2363	4185	12009	297	26	712	489	123
$E(\hat{Y})-Y$	0.00	-1.00	0.00	-1.00	-1.00	0.00	0.00	-1.00	0.00	0.00
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	0.000	-0.004	0.000	-0.023	-0.008	0.000	0.000	-0.140	0.000	0.000
$E(\hat{S})$	6.00	4.00	1.00	2.00	3.00	0.00	0.00	1.00	1.00	0.00
$S(\hat{Y})$	6.00	3.87	1.41	2.00	3.16	0.00	0.00	1.00	0.00	0.00
$E(\hat{S})-S(\hat{Y})$	0.00	0.13	-0.41	0.00	-0.16	0.00	0.00	0.00	1.00	0.00
$E(\hat{V})$	35.00	14.00	2.00	3.00	9.00	0.00	0.00	1.00	0.00	0.00
$V(\hat{Y})$	36.00	15.00	2.00	4.00	10.00	0.00	0.00	1.00	0.00	0.00
$E(\hat{V})-V(\hat{Y})$	-1.00	-1.00	0.00	-1.00	-1.00	0.00	0.00	0.00	0.00	0.00
Coverage	94.80	94.10	94.30	91.80	90.90	79.20	98.20	88.10	95.50	91.40
$MSE(\hat{Y})$	6.00	4.00	1.41	2.24	3.32	0.00	0.00	1.41	0.00	0.00
$\sqrt{MSE(\hat{Y})}$	0.00	0.02	0.06	0.05	0.03	0.00	0.00	0.20	0.00	0.00
$\left(\sqrt{MSE(\hat{Y})}/Y\right)*100$	36.00	16.00	2.00	5.00	11.00	0.00	0.00	2.00	0.00	0.00
$\left(MSE(\hat{Y})/Y\right)*100$	0.03	0.07	0.08	0.12	0.09	0.00	0.00	0.28	0.00	0.00

Appendix 3

Estimator Properties and Performance Indicators
Frequency Distribution Method

	Country of Birth - Ethnic									
	Categories									
	11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Y	2923	525	1730	16674	1509	1102	1597	782	226	3270
$E(\hat{Y})$	2922	525	1730	16673	1509	1101	1597	781	226	3277
$E(\hat{Y})-Y$	-1.00	0.00	0.00	-1.00	0.00	-1.00	0.00	-1.00	0.00	7.00
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	-0.034	0.000	0.000	-0.006	0.000	-0.090	0.000	-0.127	0.000	0.214
$E(\hat{S})$	2.00	1.00	1.00	3.00	1.00	1.00	1.00	1.00	0.00	2.00
$S(\hat{Y})$	1.73	0.00	1.00	3.46	1.00	1.00	1.00	1.00	0.00	2.65
$E(\hat{S})-S(\hat{Y})$	0.27	1.00	0.00	-0.46	0.00	0.00	0.00	0.00	0.00	-0.65
$E(\hat{V})$	2.00	0.00	1.00	12.00	1.00	1.00	1.00	1.00	0.00	3.00
$V(\hat{Y})$	3.00	0.00	1.00	12.00	1.00	1.00	1.00	1.00	0.00	7.00
$E(\hat{V})-V(\hat{Y})$	-1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-4.00
Coverage	88.90	95.30	95.00	94.10	96.30	79.30	95.70	88.40	82.80	6.00
$MSE(\hat{Y})$	2.00	0.00	1.00	3.61	1.00	1.41	1.00	1.41	0.00	7.48
$\sqrt{MSE(\hat{Y})}$	0.07	0.00	0.06	0.02	0.07	0.13	0.06	0.18	0.00	0.23
$\left(\sqrt{MSE(\hat{Y})}/Y\right)*100$	4.00	0.00	1.00	13.00	1.00	2.00	1.00	2.00	0.00	56.00
$\left(MSE(\hat{Y})/Y\right)*100$	0.14	0.00	0.06	0.08	0.07	0.18	0.06	0.26	0.00	1.71

Appendix 3

Estimator Properties and Performance Indicators
Frequency Distribution Method

	Ethnic - Long Term Illness							
	Categories							
	1	2	3	4	5	6	7	8
Y	17921	124268	3744	35980	740	6307	812	8984
$E(\hat{Y})$	17920	124268	3743	35979	739	6307	811	8989
$E(\hat{Y})-Y$	-1.00	0.00	-1.00	-1.00	-1.00	0.00	-1.00	5.00
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	-0.005	0.000	-0.026	-0.002	-0.135	0.000	-0.123	0.055
$E(\hat{S})$	5.00	9.00	3.00	7.00	1.00	3.00	1.00	4.00
$S(\hat{Y})$	5.57	8.89	2.83	7.00	1.41	3.32	1.41	4.58
$E(\hat{S})-S(\hat{Y})$	-0.57	0.11	0.17	0.00	-0.41	-0.32	-0.41	-0.58
$E(\hat{V})$	29.00	78.00	7.00	51.00	2.00	11.00	2.00	17.00
$V(\hat{Y})$	31.00	79.00	8.00	49.00	2.00	11.00	2.00	21.00
$E(\hat{V})-V(\hat{Y})$	-2.00	-1.00	-1.00	2.00	0.00	0.00	0.00	-4.00
Coverage	94.00	95.00	93.90	95.30	91.80	93.80	86.00	77.00
$MSE(\hat{Y})$	5.66	8.89	3.00	7.07	1.73	3.32	1.73	6.78
$\sqrt{MSE(\hat{Y})}$	0.03	0.01	0.08	0.02	0.23	0.05	0.21	0.08
$\left(\sqrt{MSE(\hat{Y})}/Y\right)*100$	32.00	79.00	9.00	50.00	3.00	11.00	3.00	46.00
$\left(MSE(\hat{Y})/Y\right)*100$	0.18	0.06	0.24	0.14	0.41	0.17	0.37	0.51

Appendix 3

Estimator Properties and Performance Indicators
Frequency Distribution Method

	Country of Birth - Long Term Illness									
	Categories									
	1	2	3	4	5	6	7	8	9	10
Y	17553	137207	1454	11592	694	3366	3015	18000	501	5374
$E(\hat{Y})$	17553	137207	1452	11591	694	3365	3014	17999	500	5380
$E(\hat{Y})-Y$	0.00	0.00	-2.00	-1.00	0.00	-1.00	-1.00	-1.00	-1.00	6.00
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	0.000	0.000	-0.137	-0.008	0.000	-0.029	-0.033	-0.005	-0.199	0.111
$E(\hat{S})$	4.00	6.00	1.00	3.00	1.00	2.00	2.00	3.00	1.00	2.00
$S(\hat{Y})$	3.74	5.92	1.41	3.61	1.00	1.73	1.73	3.46	1.00	2.65
$E(\hat{S})-S(\hat{Y})$	0.26	0.08	-0.41	-0.61	0.00	0.27	0.27	-0.46	0.00	-0.65
$E(\hat{V})$	14.00	34.00	2.00	12.00	1.00	3.00	3.00	11.00	1.00	5.00
$V(\hat{Y})$	14.00	35.00	2.00	13.00	1.00	3.00	3.00	12.00	1.00	7.00
$E(\hat{V})-V(\hat{Y})$	0.00	-1.00	0.00	-1.00	0.00	0.00	0.00	-1.00	0.00	-2.00
Coverage	95.00	94.70	79.00	93.70	90.80	91.40	91.50	92.60	88.60	24.70
$MSE(\hat{Y})$	3.74	5.92	2.45	3.74	1.00	2.00	2.00	3.61	1.41	6.56
$\sqrt{MSE(\hat{Y})}$	0.02	0.00	0.17	0.03	0.14	0.06	0.07	0.02	0.28	0.12
$\left(\sqrt{MSE(\hat{Y})}/Y\right)*100$	14.00	35.00	6.00	14.00	1.00	4.00	4.00	13.00	2.00	43.00
$\left(MSE(\hat{Y})/Y\right)*100$	0.08	0.03	0.41	0.12	0.14	0.12	0.13	0.07	0.40	0.80

Appendix 3

Estimator Properties and Performance Indicators
Highest Probability Method

	Country of Birth - Ethnic									
	Categories									
	1	2	3	4	5	6	7	8	9	10
Y	126363	21848	2363	4186	12010	297	26	713	489	123
$E(\hat{Y})$	126420	21830	2361	4183	12001	297	26	712	489	123
$E(\hat{Y})-Y$	57.00	-18.00	-2.00	-3.00	-9.00	0.00	0.00	-1.00	0.00	0.00
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	0.045	-0.082	-0.084	-0.071	-0.074	0.000	0.000	-0.140	0.000	0.000
$S(\hat{Y})$	6.08	4.00	1.41	1.73	3.16	0.00	0.00	1.00	0.00	0.00
$V(\hat{Y})$	37.00	16.00	2.00	3.00	10.00	0.00	0.00	1.00	0.00	0.00
Coverage	0	0	71.3	58.3	18.7	79.3	98.2	88.1	68.7	91.4
$MSE(\hat{Y})$	57.32	18.44	2.45	3.46	9.54	0.00	0.00	1.41	0.00	0.00
$\sqrt{MSE(\hat{Y})}$	0.05	0.08	0.10	0.08	0.08	0.00	0.00	0.20	0.00	0.00
$\left(\sqrt{MSE(\hat{Y})}/Y\right)*100$	3286.00	340.00	6.00	12.00	91.00	0.00	0.00	2.00	0.00	0.00
$\left(MSE(\hat{Y})/Y\right)*100$	2.60	1.56	0.25	0.29	0.76	0.00	0.00	0.28	0.00	0.00

Appendix 3

Estimator Properties and Performance Indicators

Highest Probability Method

	Country of Birth - Ethnic									
	Categories									
	11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Y	2923	525	1730	16674	1509	1102	1597	782	226	3270
$E(\hat{Y})$	2921	525	1729	16661	1508	1101	1596	781	226	3267
$E(\hat{Y})-Y$	-2.00	0.00	-1.00	-13.00	-1.00	-1.00	-1.00	-1.00	0.00	-3.00
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	-0.068	0.000	-0.057	-0.077	-0.066	-0.090	-0.062	-0.127	0.000	-0.091
$S(\hat{Y})$	1.41	0.00	1.00	3.46	1.00	1.00	1.00	1.00	0.00	1.73
$V(\hat{Y})$	2.00	0.00	1.00	12.00	1.00	1.00	1.00	1.00	0.00	3.00
Coverage	59.1	65.8	59.7	2.7	64.3	78.7	66.4	88.4	82.8	74.8
$MSE(\hat{Y})$	2.45	0.00	1.41	13.45	1.41	1.41	1.41	1.41	0.00	3.46
$\sqrt{MSE(\hat{Y})}$	0.08	0.00	0.08	0.08	0.09	0.13	0.09	0.18	0.00	0.11
$\left(\sqrt{MSE(\hat{Y})}/Y\right)*100$	6.00	0.00	2.00	181.00	2.00	2.00	2.00	2.00	0.00	12.00
$\left(MSE(\hat{Y})/Y\right)*100$	0.21	0.00	0.12	1.09	0.13	0.18	0.13	0.26	0.00	0.37

Appendix 3

Estimator Properties and Performance Indicators
Highest Probability Method

	Ethnic - Long Term Illness							
	Categories							
	1	2	3	4	5	6	7	8
Y	17921	124268	3744	35980	740	6307	812	8984
$E(\hat{Y})$	17885	124364	3738	35941	738	6302	810	8977
$E(\hat{Y})-Y$	-36.00	96.00	-6.00	-39.00	-2.00	-5.00	-2.00	-7.00
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	-0.200	0.077	-0.160	-0.108	-0.270	-0.079	-0.246	-0.077
$S(\hat{Y})$	6.40	10.00	3.00	8.06	1.41	3.46	1.41	4.58
$V(\hat{Y})$	41.00	100.00	9.00	65.00	2.00	12.00	2.00	21.00
Coverage	0	0	51.9	0.3	77.1	63.9	71.7	61.4
$MSE(\hat{Y})$	36.57	96.52	6.71	39.82	2.45	6.08	2.45	8.37
$\sqrt{MSE(\hat{Y})}$	0.20	0.08	0.18	0.11	0.33	0.10	0.30	0.09
$\left(\sqrt{MSE(\hat{Y})}/Y\right)*100$	1337.00	9316.00	45.00	1586.00	6.00	37.00	6.00	70.00
$\left(MSE(\hat{Y})/Y\right)*100$	7.46	7.50	1.20	4.41	0.81	0.59	0.74	0.78

Appendix 3

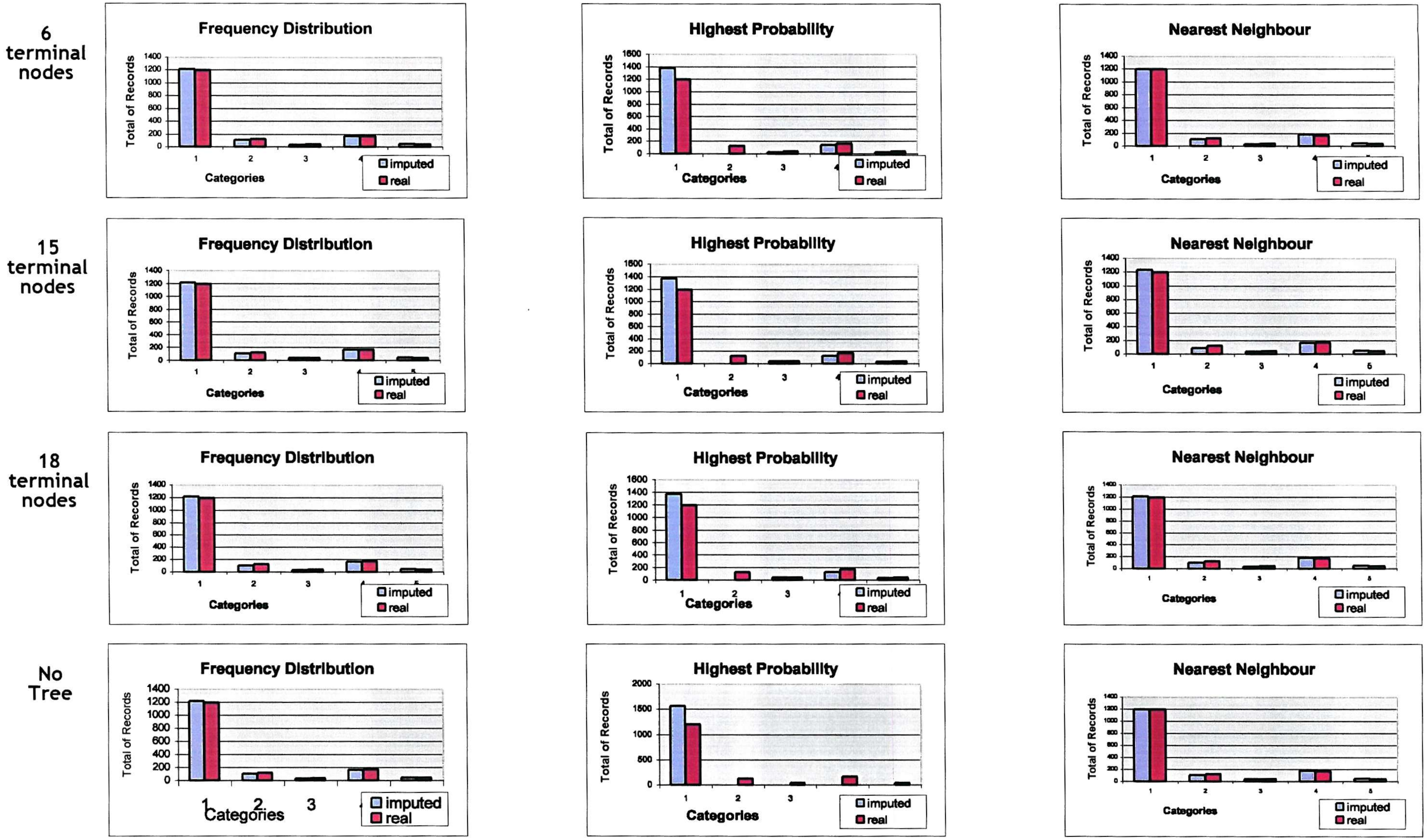
Estimator Properties and Performance Indicators

Highest Probability Method

	Country of Birth - Long Term Illness									
	Categories									
	1	2	3	4	5	6	7	8	9	10
Y	17553	137207	1454	11592	694	3366	3015	18000	501	5374
$E(\hat{Y})$	17537	137247	1452	11579	693	3366	3012	17997	500	5373
$E(\hat{Y})-Y$	-16.00	40.00	-2.00	-13.00	-1.00	0.00	-3.00	-3.00	-1.00	-1.00
$\left(\left(E(\hat{Y})-Y\right)/Y\right)*100$	-0.091	0.029	-0.137	-0.112	-0.144	0.000	-0.099	-0.016	-0.199	-0.018
$S(\hat{Y})$	4.24	6.40	1.41	3.61	1.00	2.00	1.73	3.61	1.00	2.45
$V(\hat{Y})$	18.00	41.00	2.00	13.00	1.00	4.00	3.00	13.00	1.00	6.00
Coverage	2.5	0	77.6	5.8	82.5	91.8	54.7	88.2	88.2	90.3
$MSE(\hat{Y})$	16.55	40.51	2.45	13.49	1.41	2.00	3.46	4.69	1.41	2.65
$\sqrt{MSE(\hat{Y})}$	0.09	0.03	0.17	0.12	0.20	0.06	0.11	0.03	0.28	0.05
$\left(\sqrt{MSE(\hat{Y})}/Y\right)*100$	274.00	1641.00	6.00	182.00	2.00	4.00	12.00	22.00	2.00	7.00
$\left(MSE(\hat{Y})/Y\right)*100$	1.56	1.20	0.41	1.57	0.29	0.12	0.40	0.12	0.40	0.13

Appendix 4

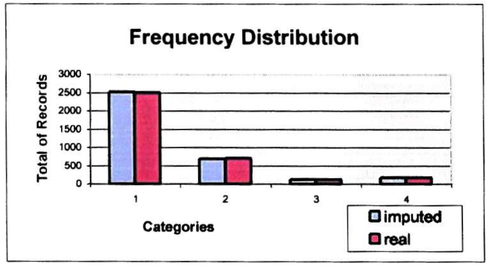
Variable: Country of Birth
Marginal Distributions

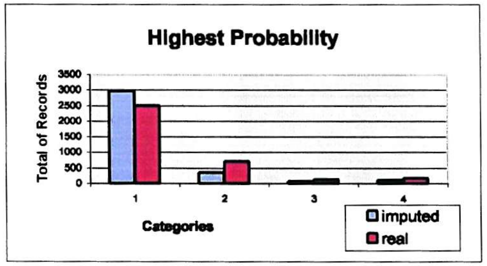


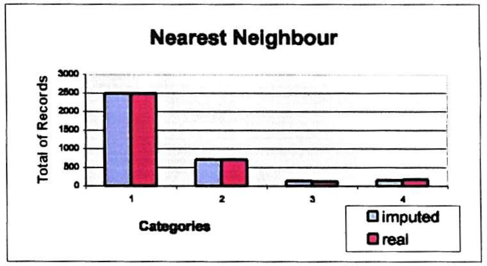
Appendix 4

Variable: Ethnic
Marginal Distributions

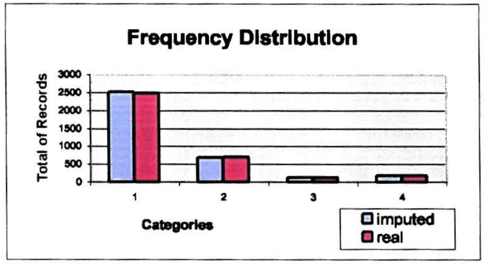
4
terminal
nodes

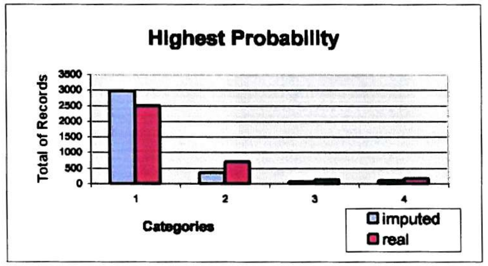


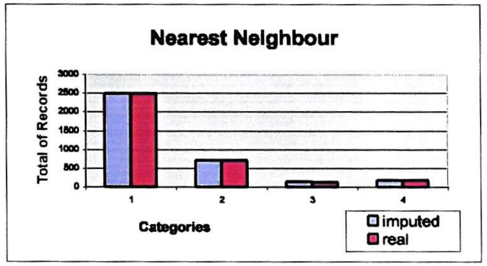




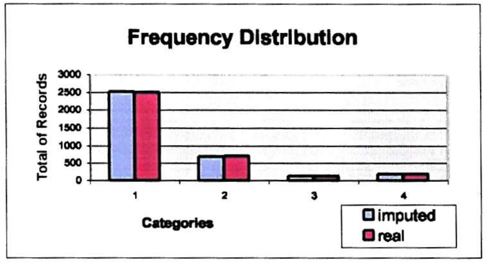
10
terminal
nodes

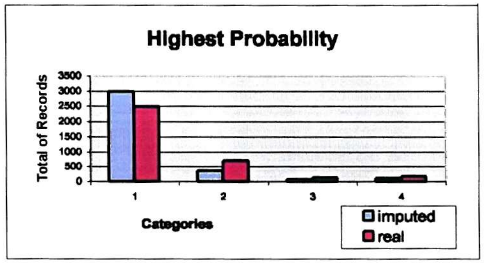


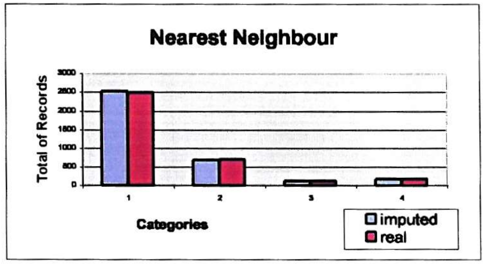




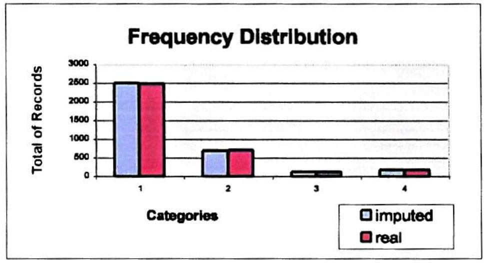
13
terminal
nodes

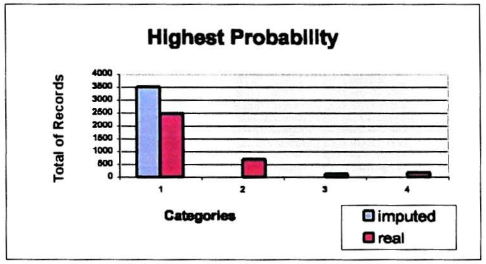


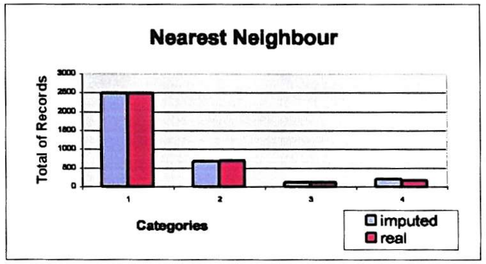




No
Tree



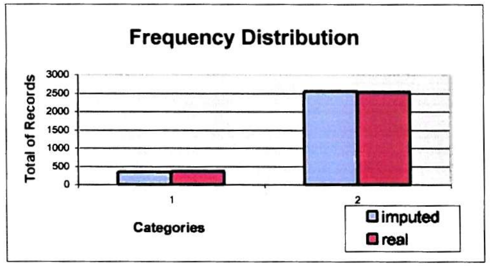


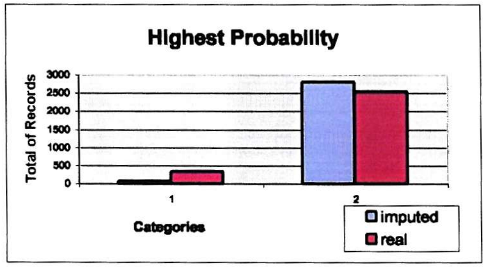


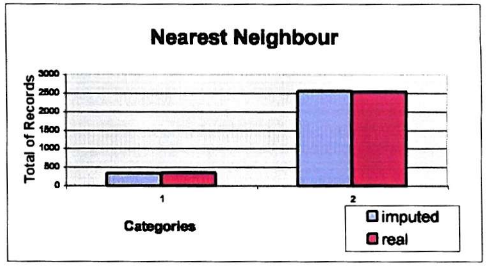
Appendix 4

Variable: Long Term Illness
Marginal Distributions

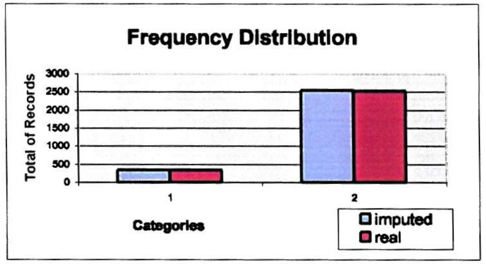
14
terminal
nodes

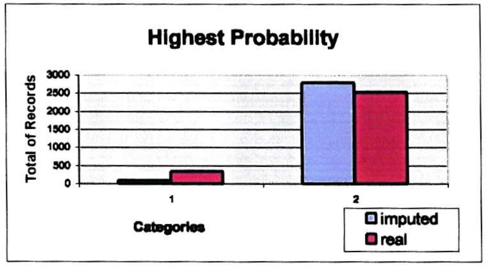


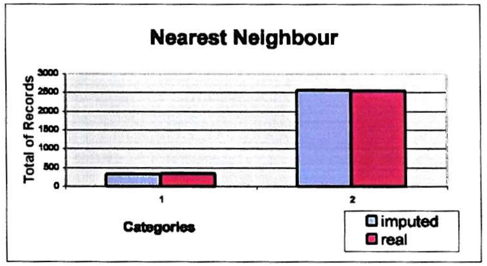




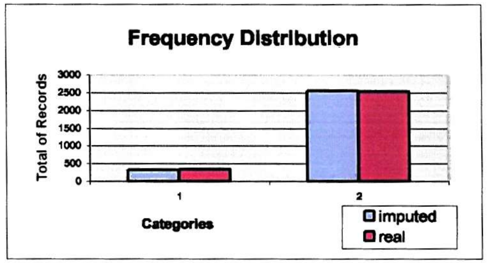
21
terminal
nodes

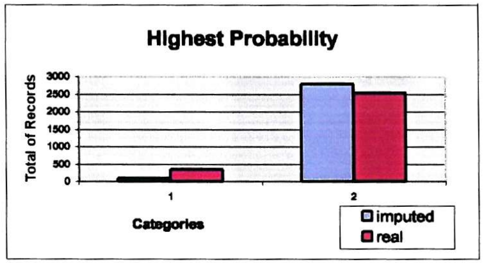


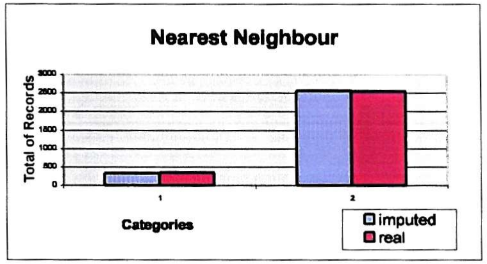




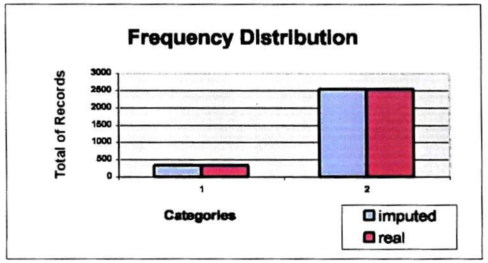
29
terminal
nodes

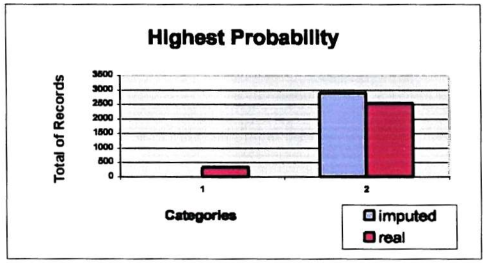


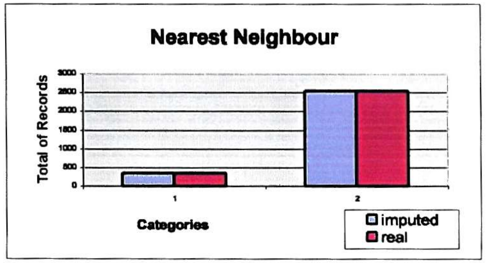




No
Tree



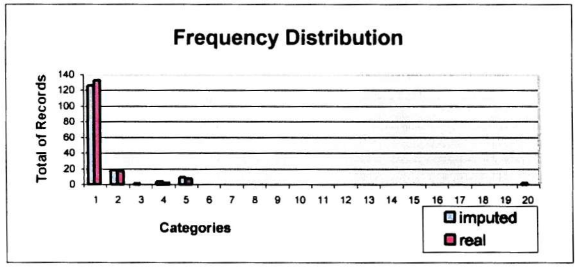


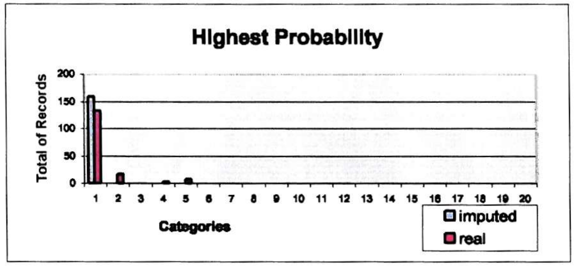


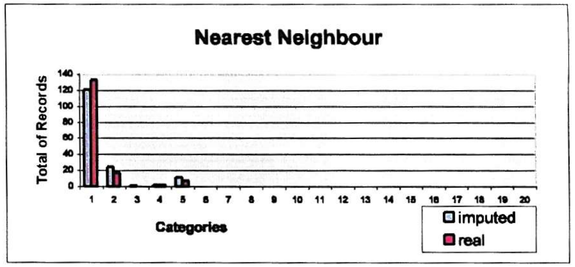
Appendix 4

Variable: Country of Birth - Ethnic
Marginal Distributions

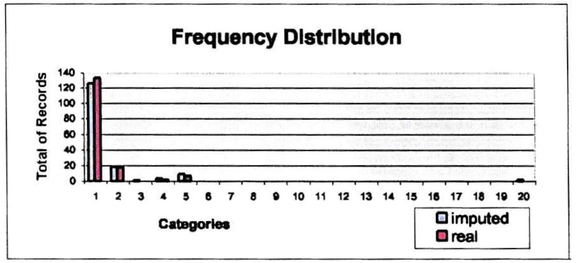
10 terminal nodes

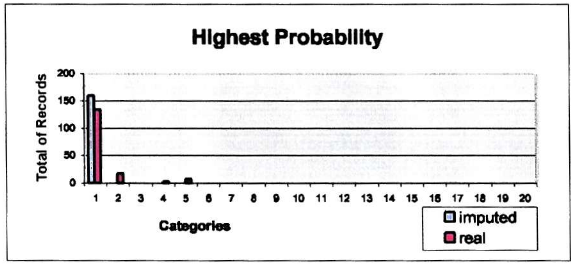


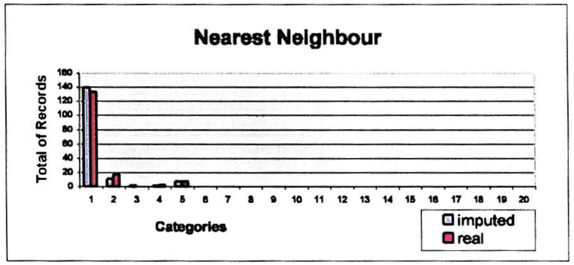




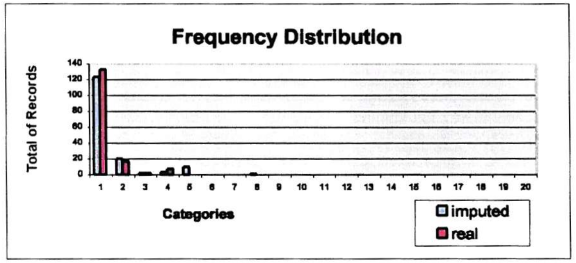
18 terminal nodes

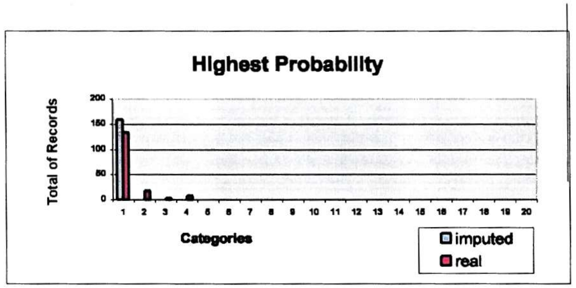


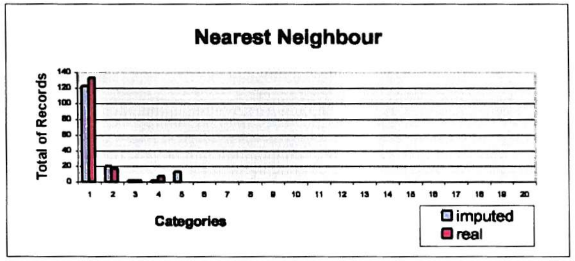




No Tree

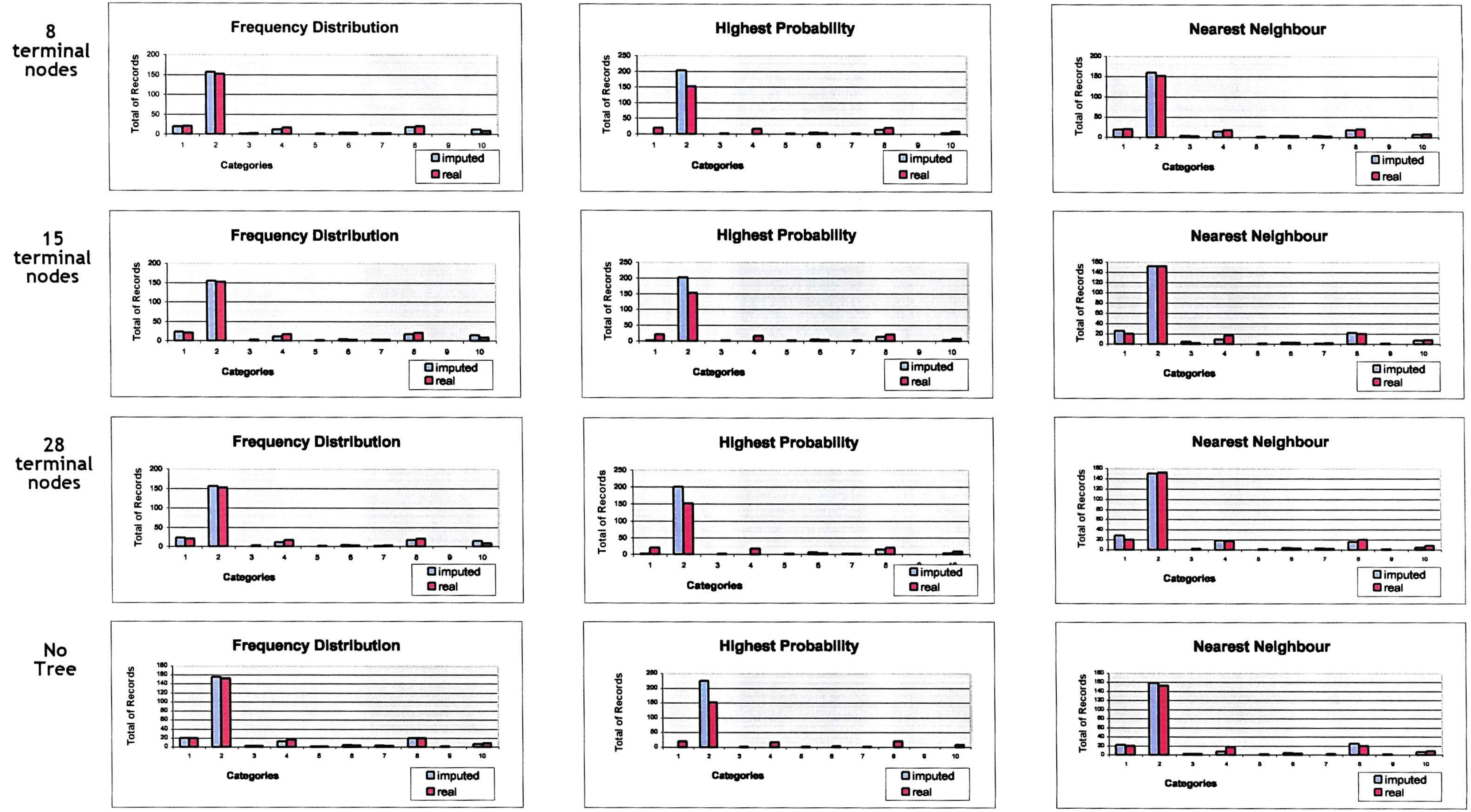






Appendix 4

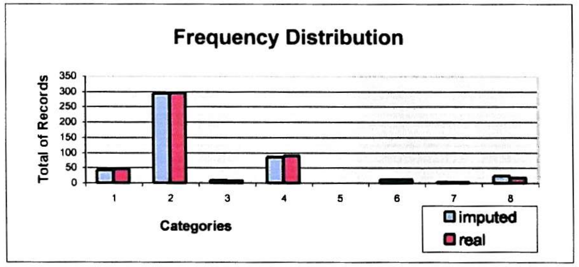
Variable: Country of Birth - Long Term Illness
Marginal Distributions

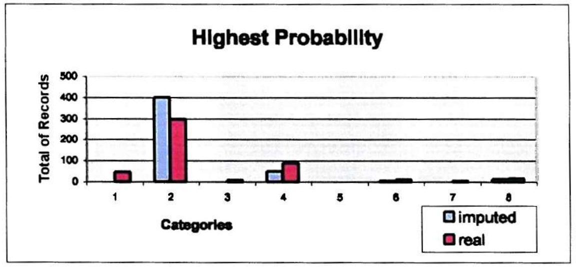


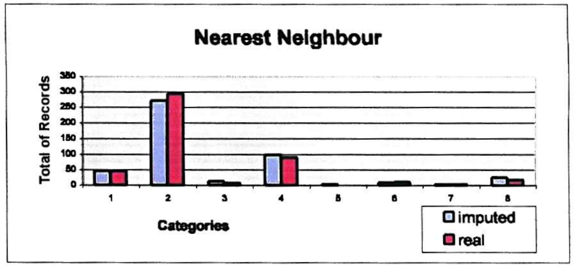
Appendix 4

Variable: Ethnic - Long Term Illness
Marginal Distributions

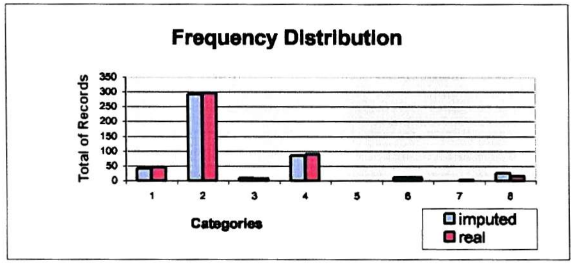
4 terminal nodes

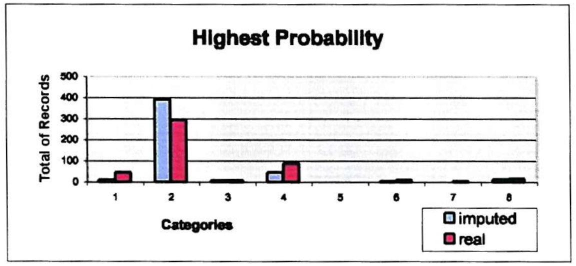


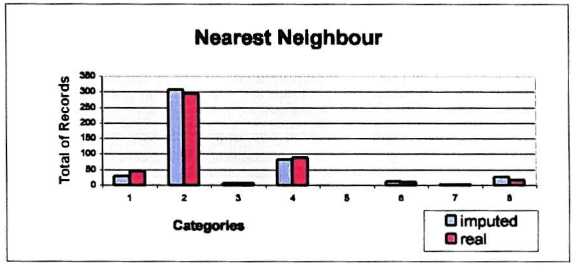




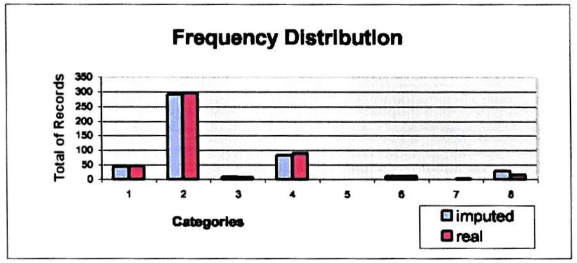
15 terminal nodes

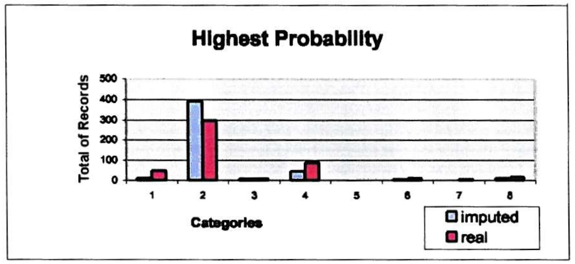


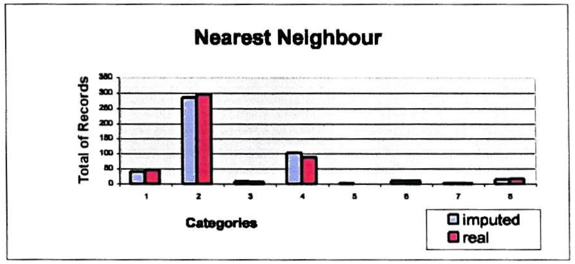




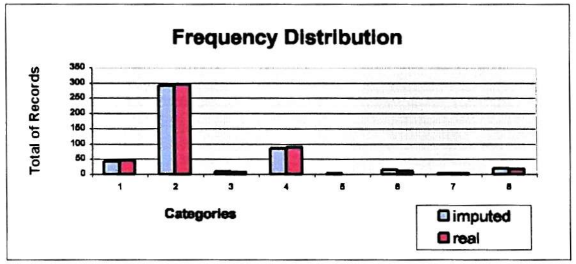
27 terminal nodes

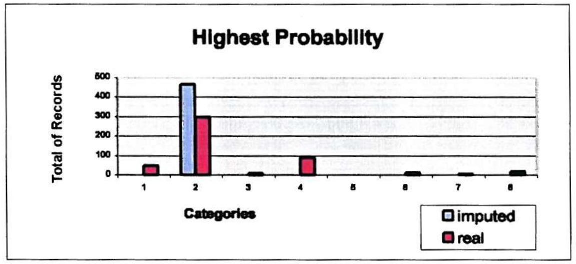


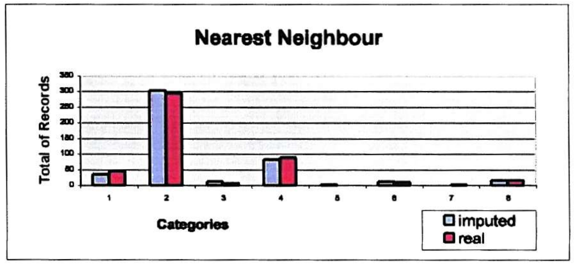




No Tree



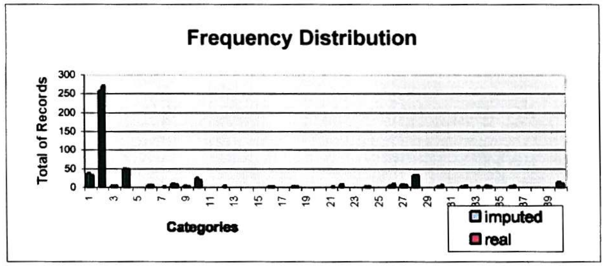


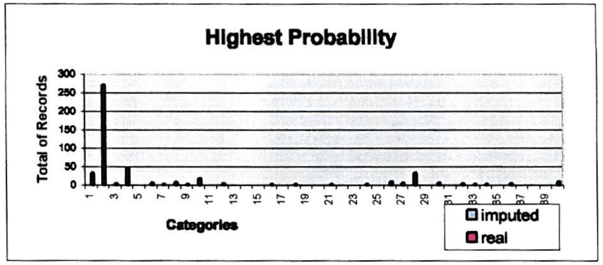


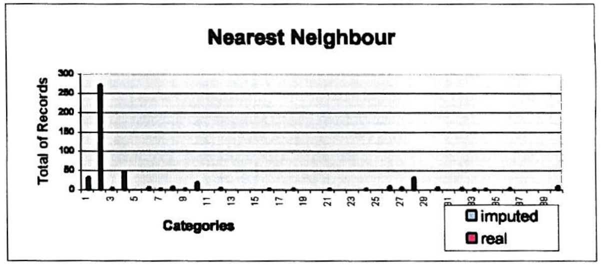
Appendix 4

Variable: Country of Birth - Ethnic - Long Term Illness
Marginal Distributions

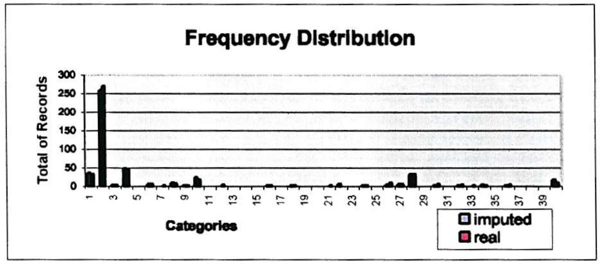
5 Terminal Nodes

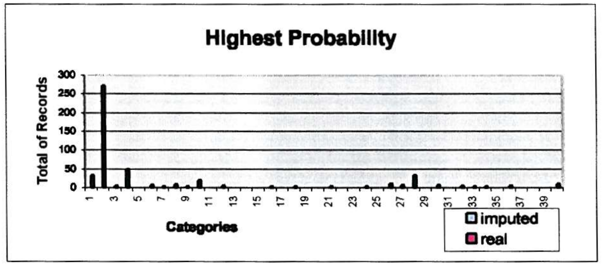


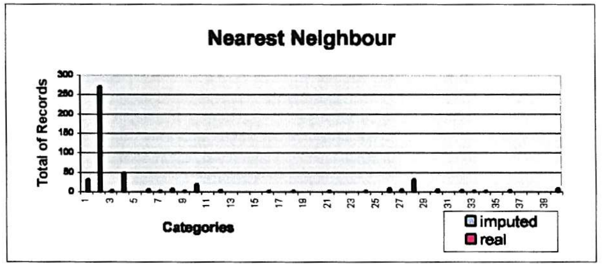




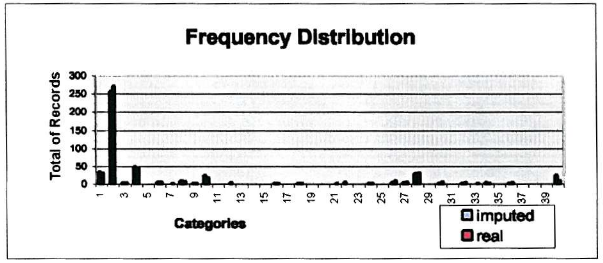
12 Terminal Nodes

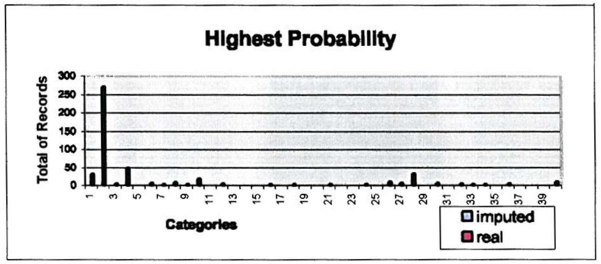


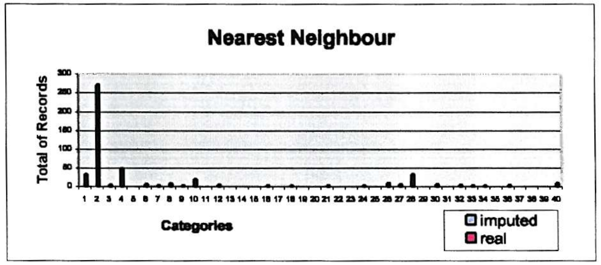




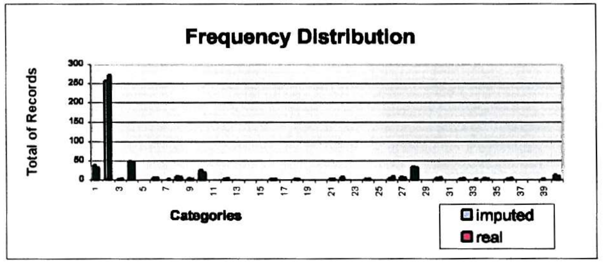
23 Terminal Nodes

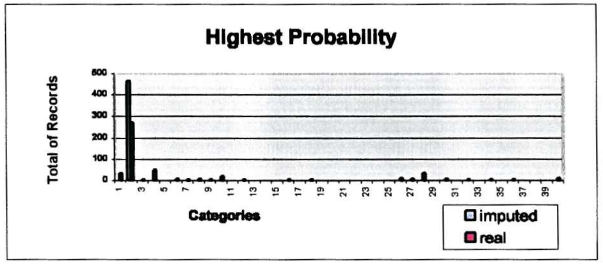


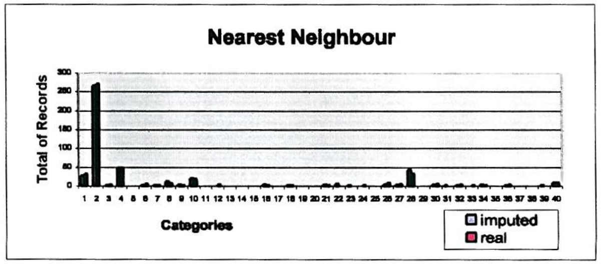




No Tree





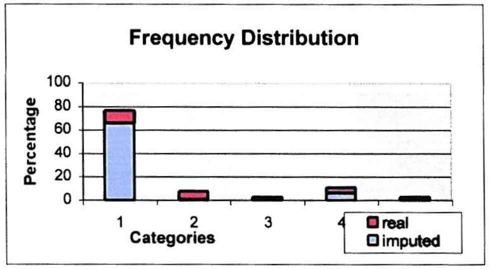


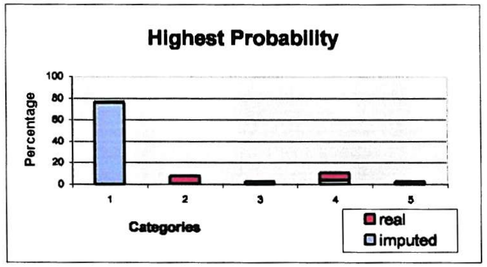
Appendix 5

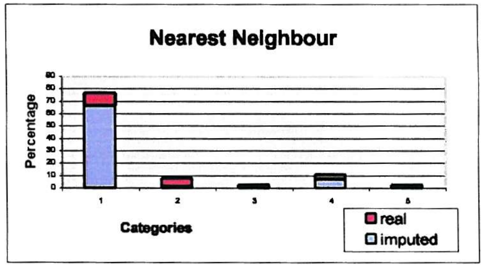
Variable: Country of Birth

Diagonals

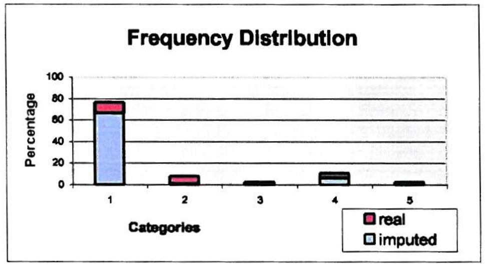
6 terminal nodes

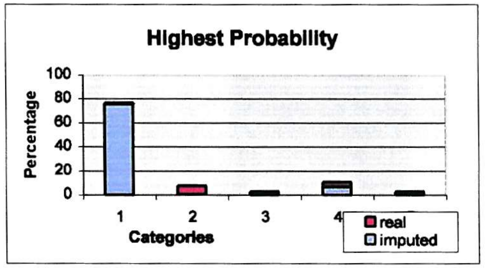


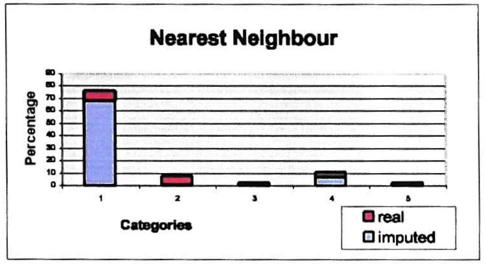




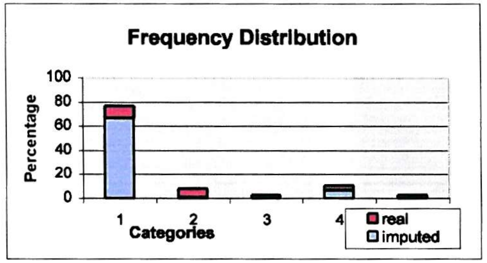
15 terminal nodes

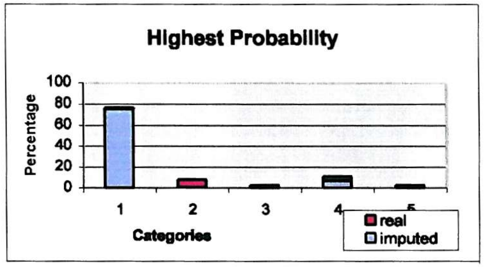


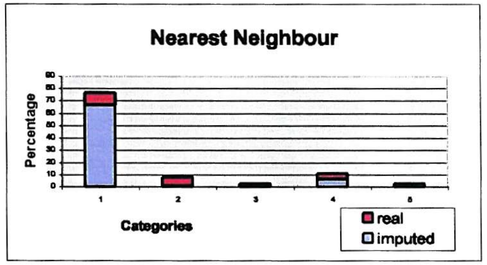




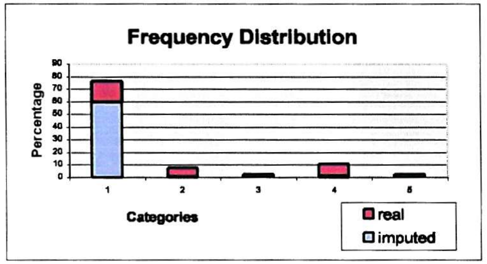
18 terminal nodes

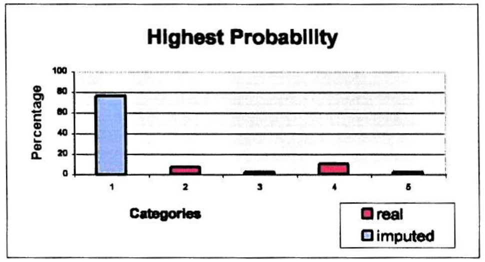


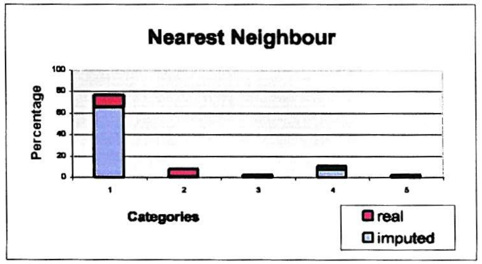




No Tree



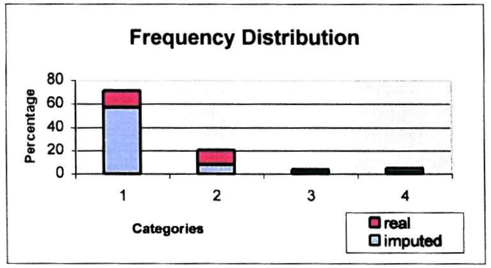


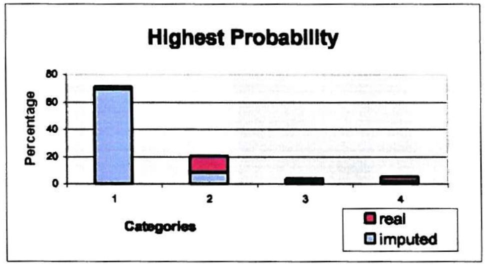


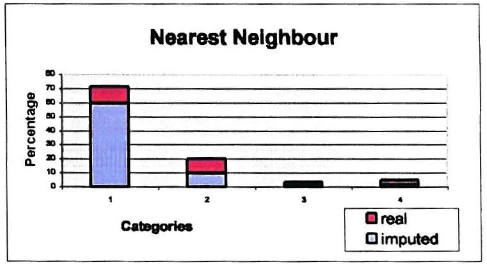
Appendix 5

Variable: Ethnic
Diagonals

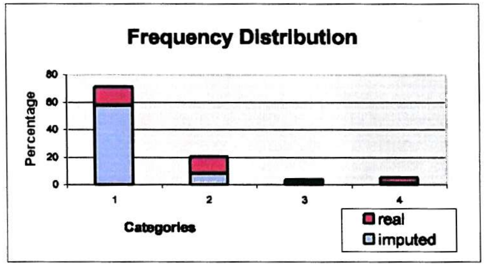
4 terminal nodes

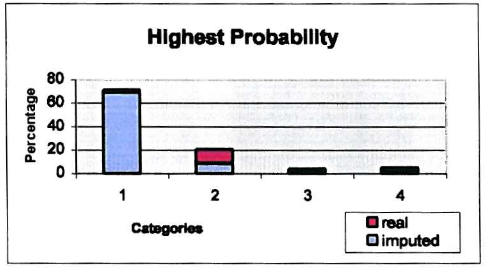


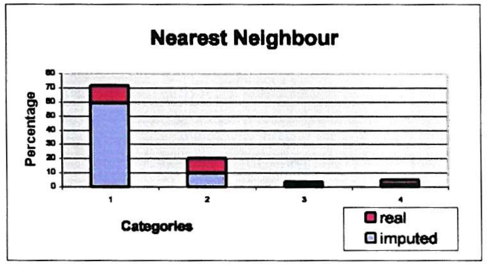




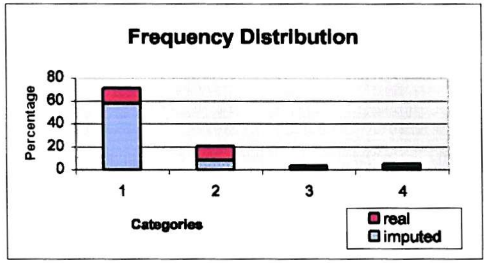
10 terminal nodes

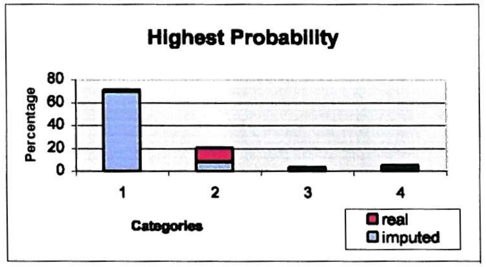


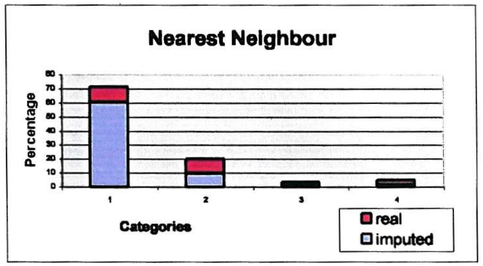




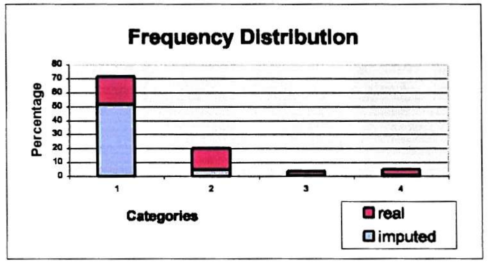
13 terminal nodes

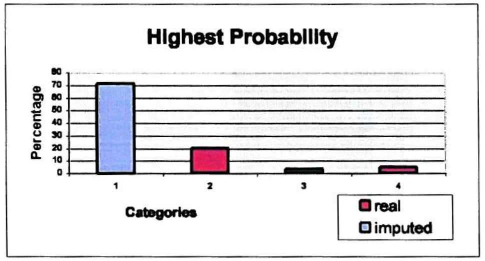


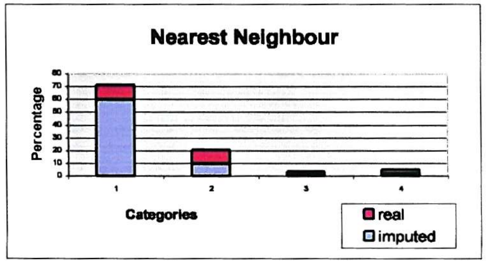




No Tree



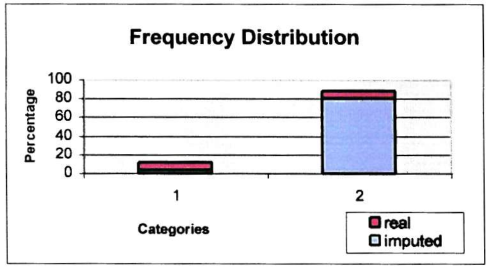


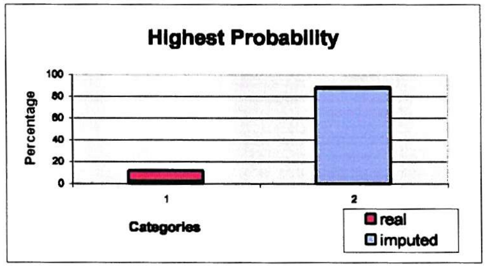


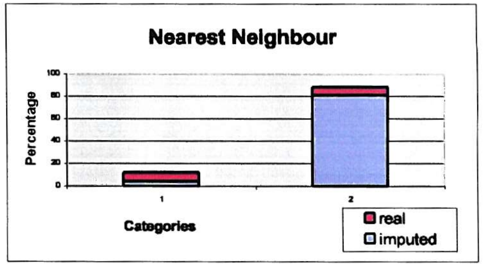
Appendix 5

Variable: Long Term Illness
Diagonals

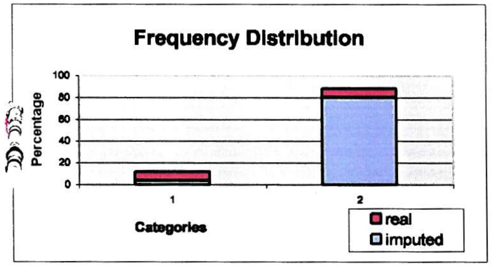
14
terminal
nodes

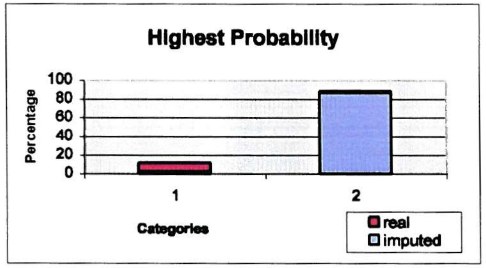


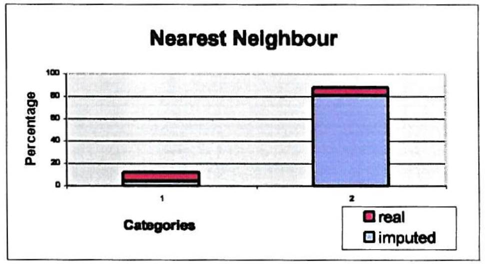




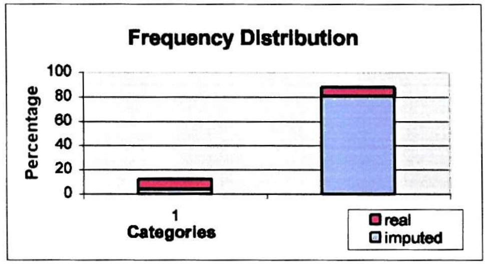
21
terminal
nodes

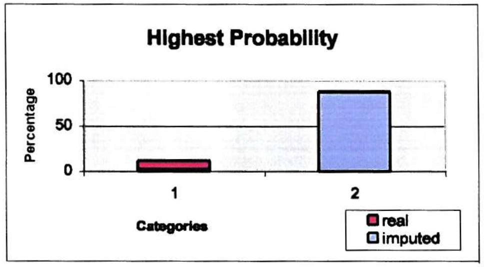


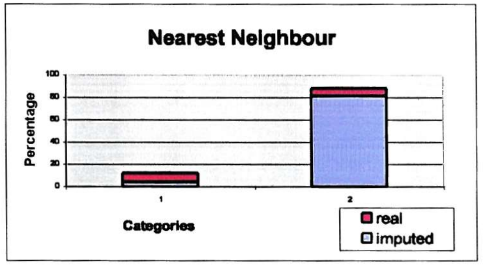




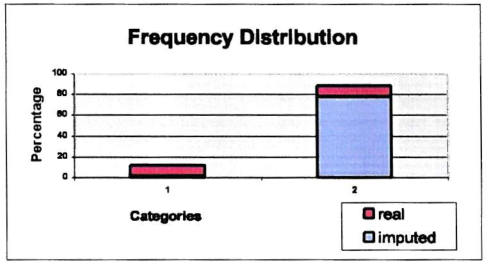
29
terminal
nodes

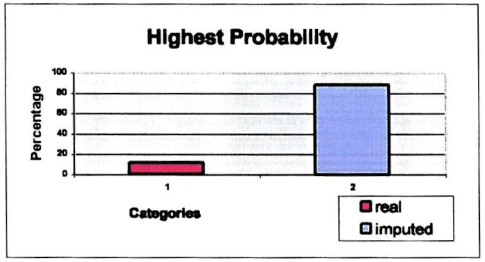


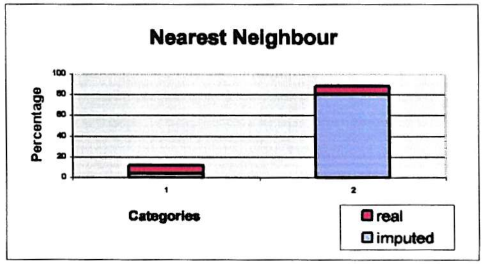




No
Tree



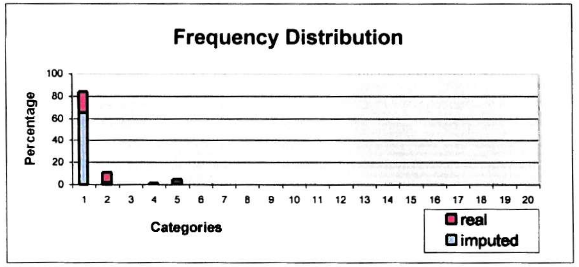


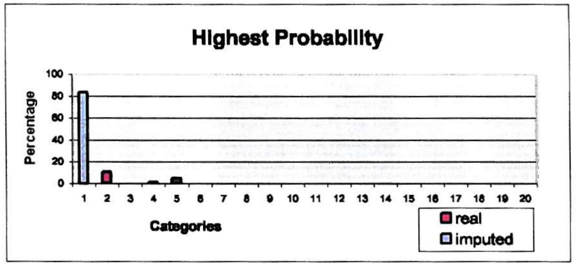


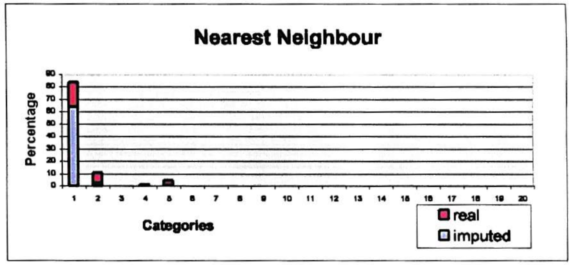
Appendix 5

Variable: Country of Birth - Ethnic Diagonals

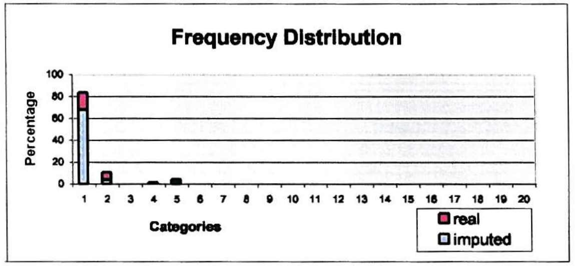
10 terminal nodes

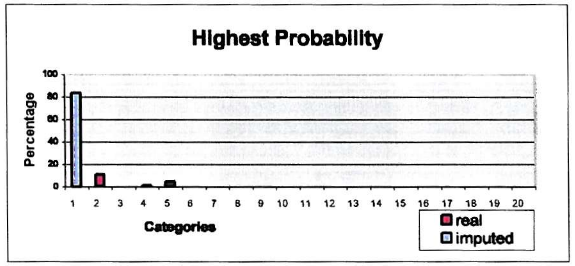


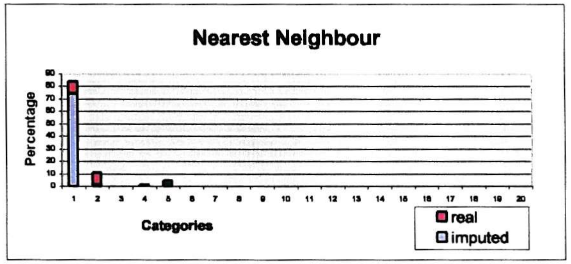




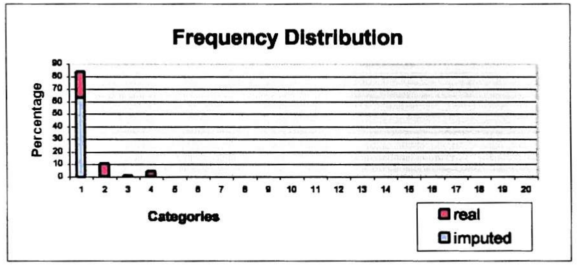
18 terminal nodes

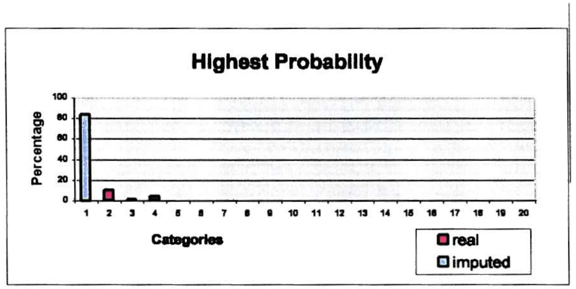


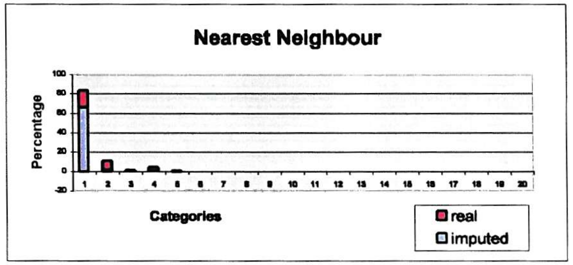




No Tree



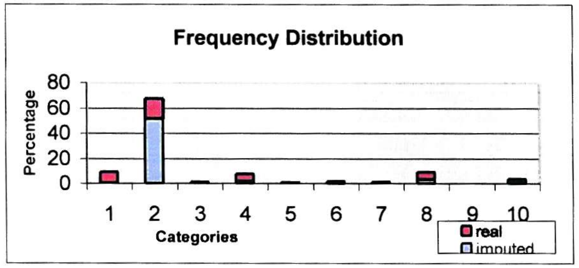


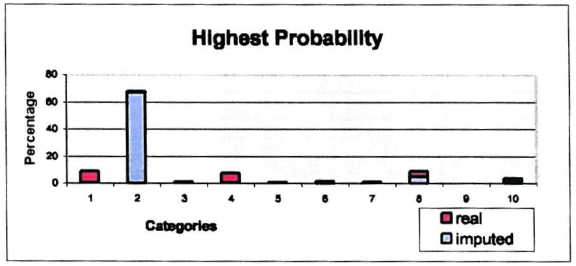


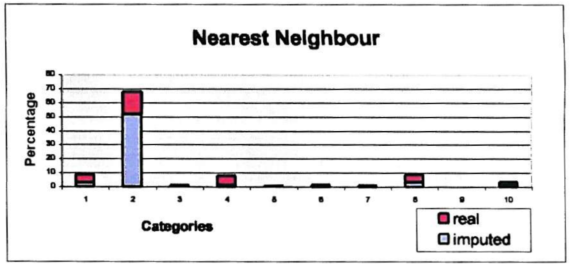
Appendix 5

Variable: Country of Birth - Long Term Illness
Diagonals

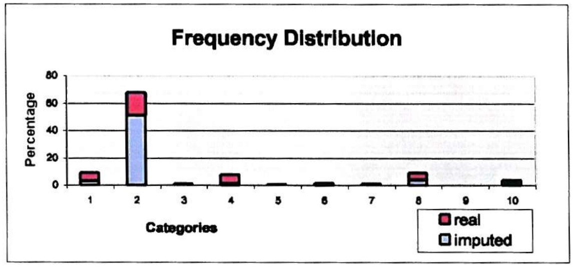
8
terminal
nodes

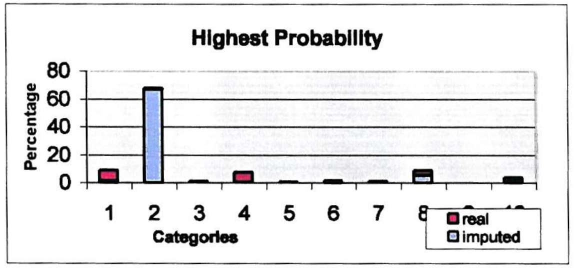


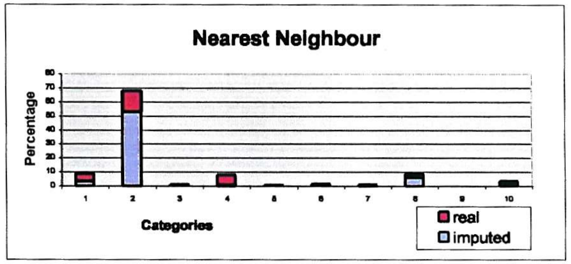




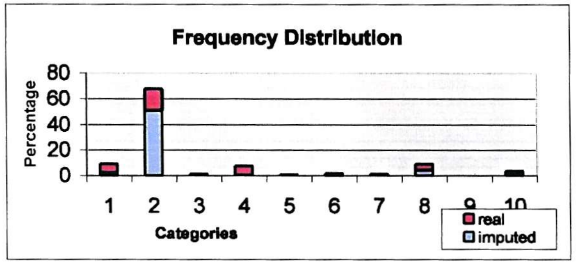
15
terminal
nodes

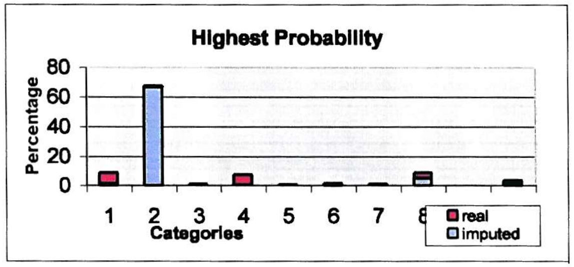


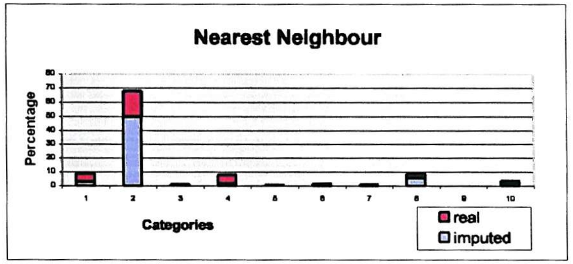




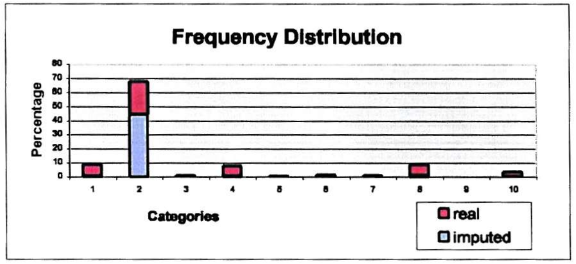
28
terminal
nodes

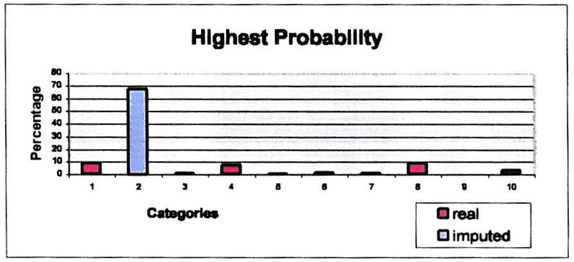


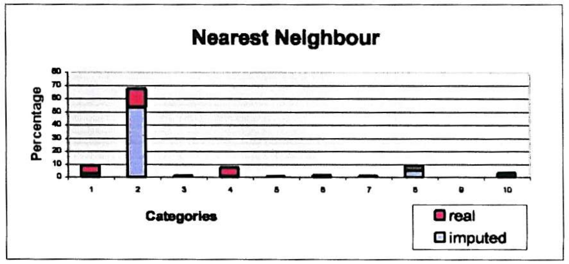




No
Tree

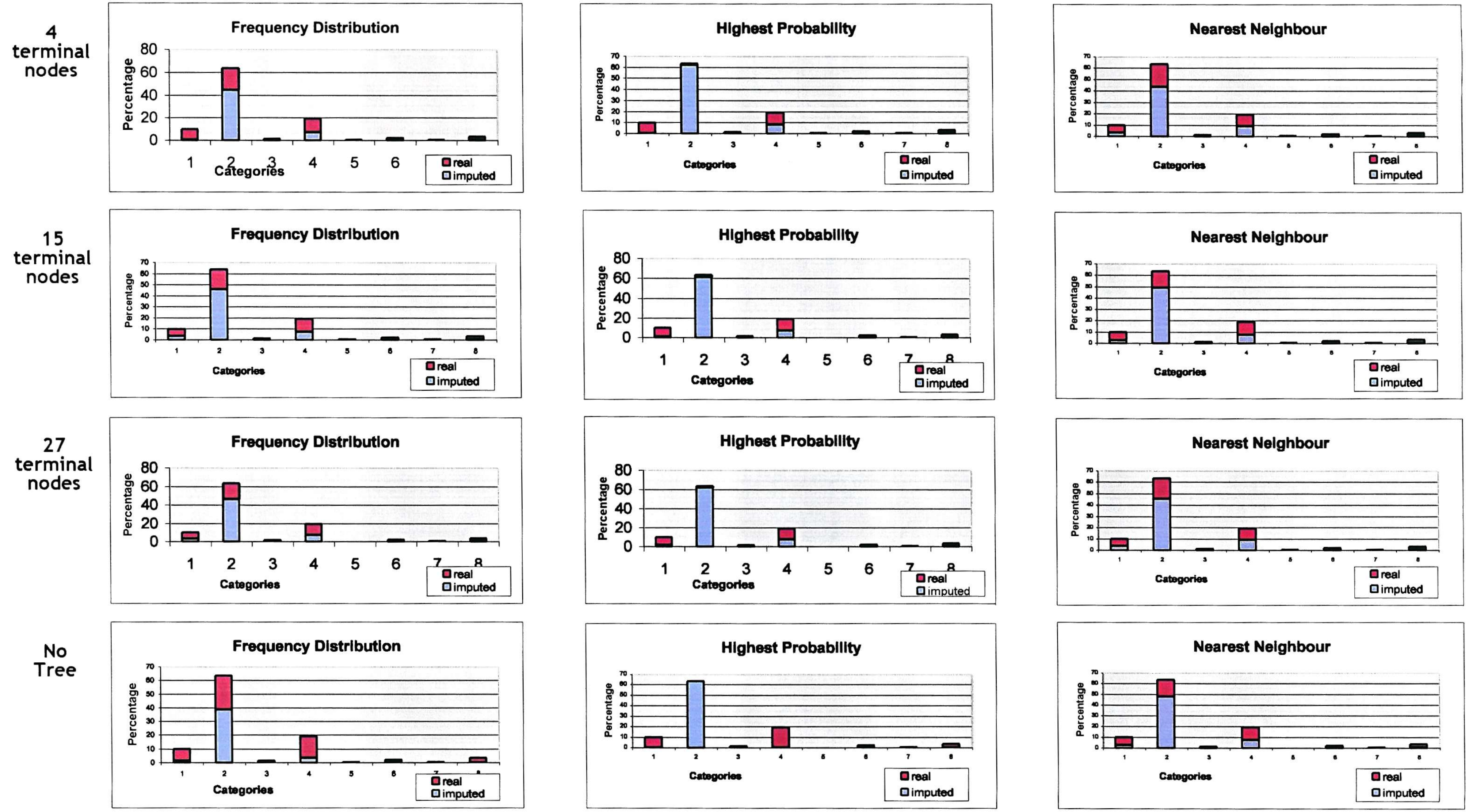






Appendix 5

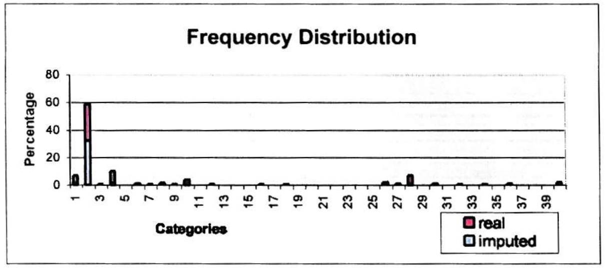
Variable: Ethnic - Long Term Illness
Diagonals

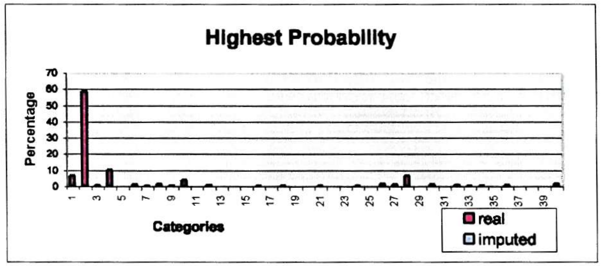


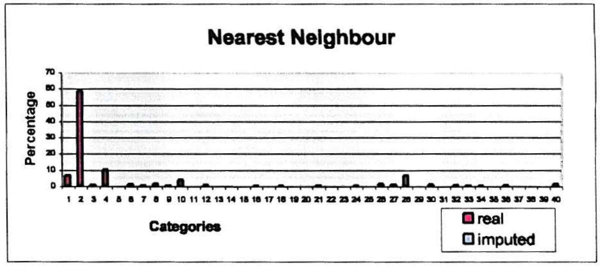
Appendix 5

Variable: Country of Birth - Ethnic - Long Term Illness
Diagonals

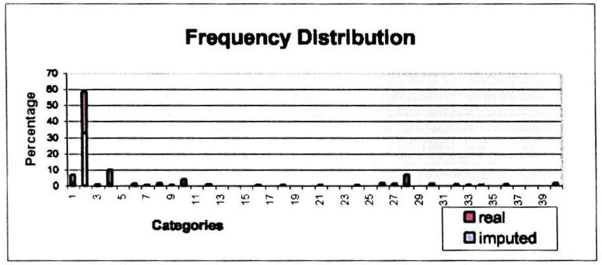
5 Terminal Nodes

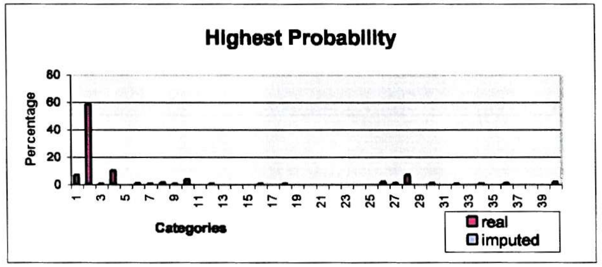

Frequency Distribution

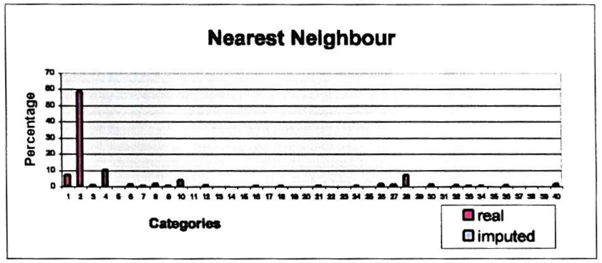

Highest Probability


Nearest Neighbour

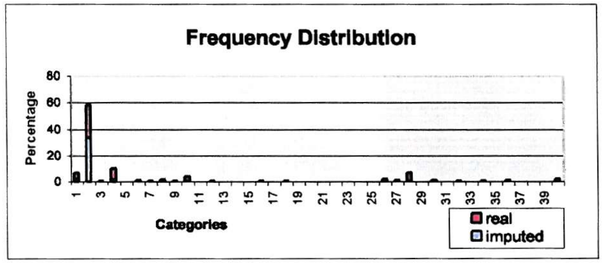
12 Terminal Nodes

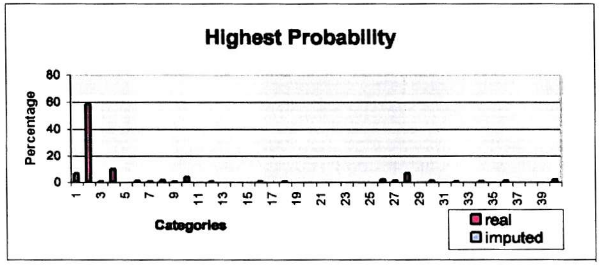

Frequency Distribution

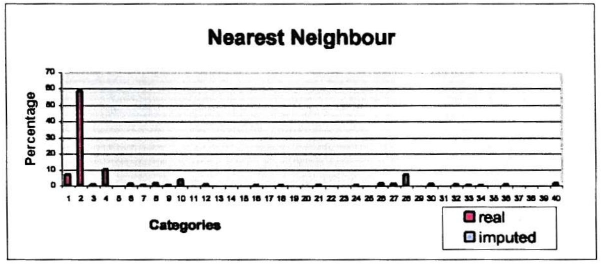

Highest Probability


Nearest Neighbour

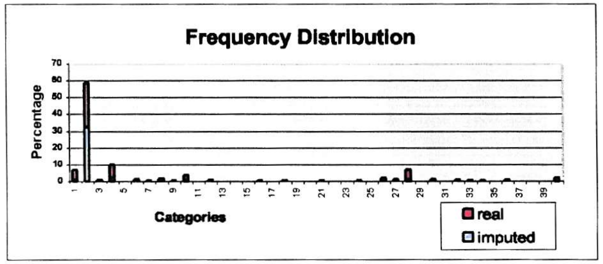
23 Terminal Nodes

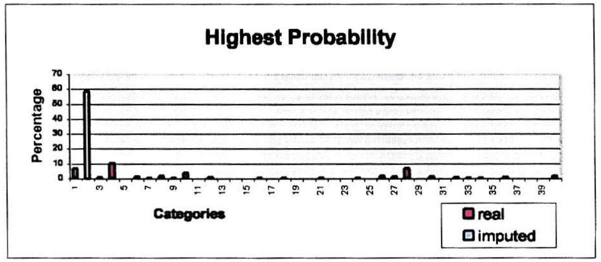

Frequency Distribution

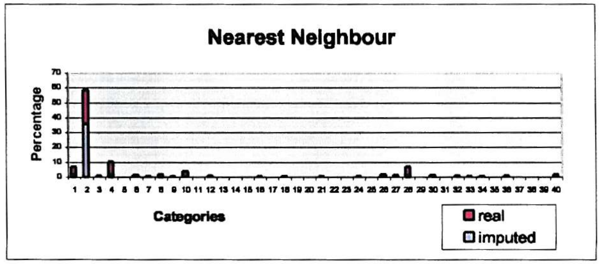

Highest Probability


Nearest Neighbour

No Tree


Frequency Distribution

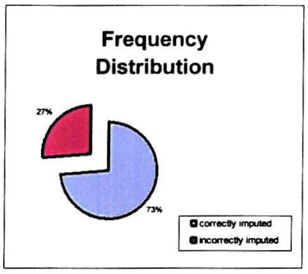

Highest Probability

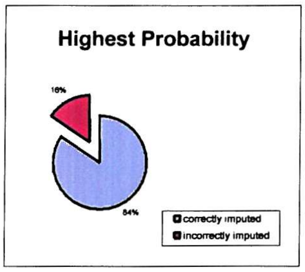

Nearest Neighbour

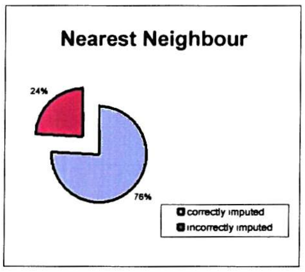
Appendix 6

Variable: Country of Birth
Accuracy of the imputation procedure

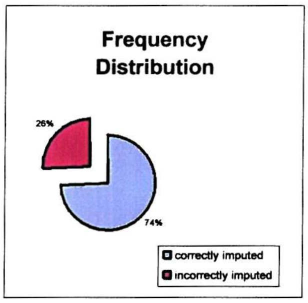
6
Terminal
Nodes

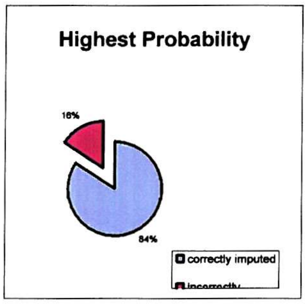


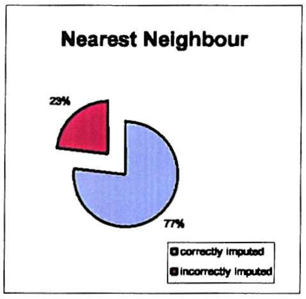




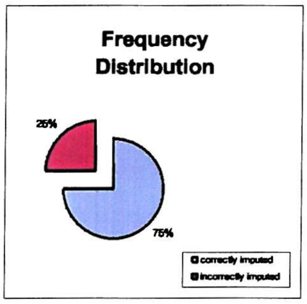
15
Terminal
Nodes

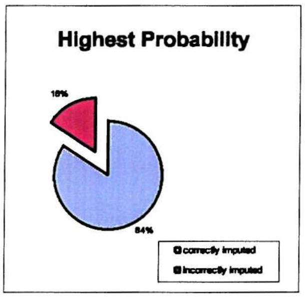


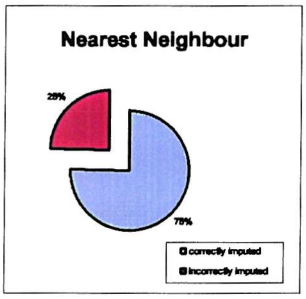




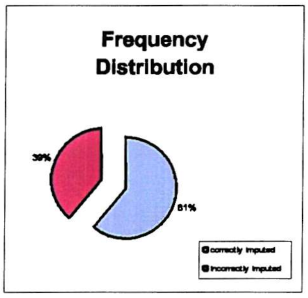
18
Terminal
Nodes

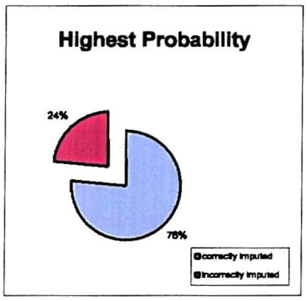


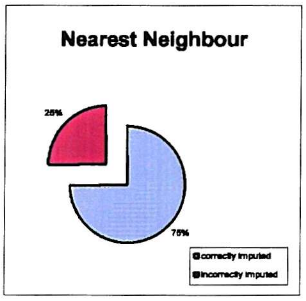




No
Tree



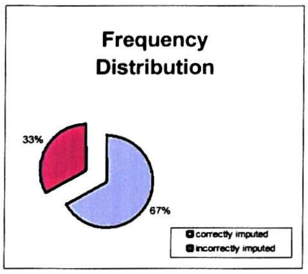


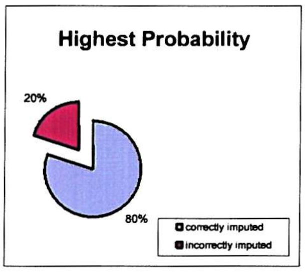


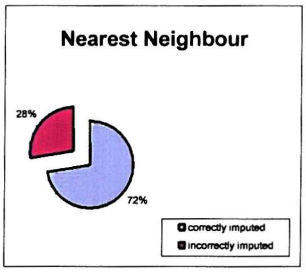
Appendix 6

Variable: Ethnic
Accuracy of the imputation procedure

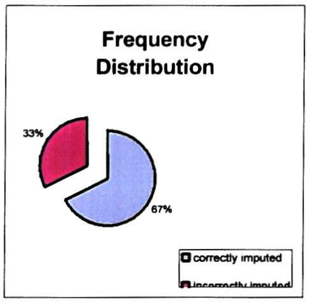
4 Terminal Nodes

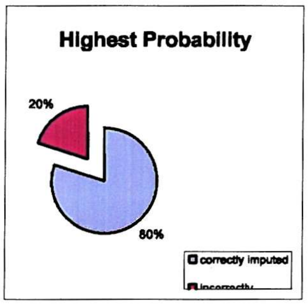


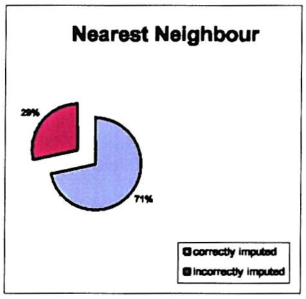




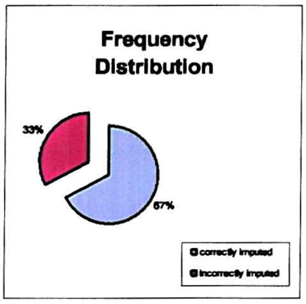
10 Terminal Nodes

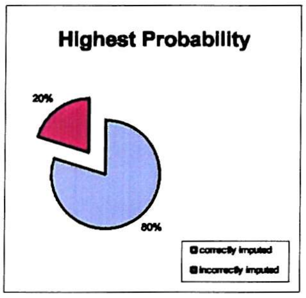


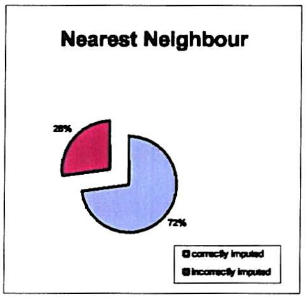




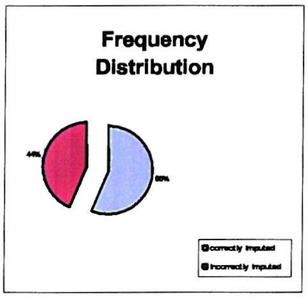
13 Terminal Nodes

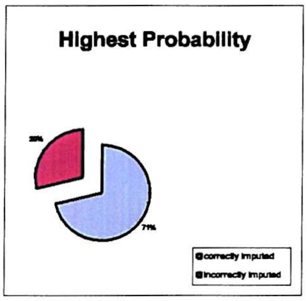


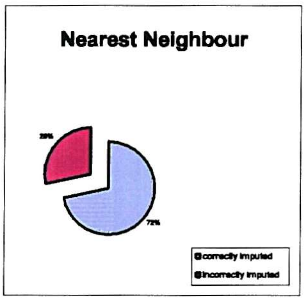




No Tree



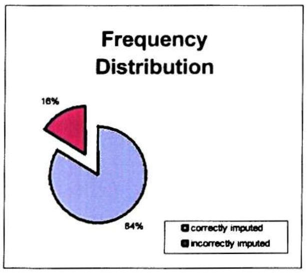


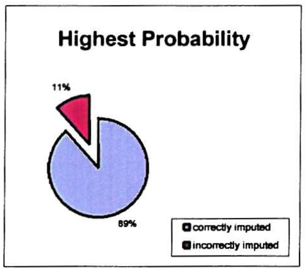


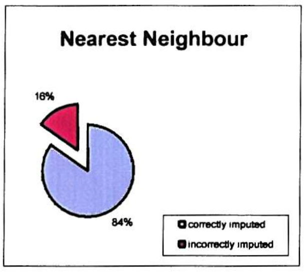
Appendix 6

Variable: Long Term Illness
Accuracy of the imputation procedure

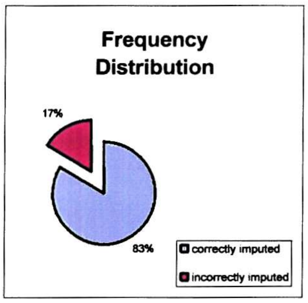
4
Terminal
Nodes

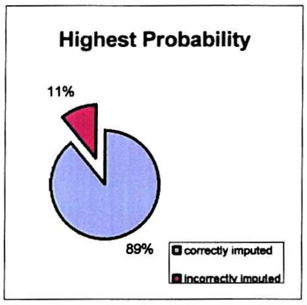


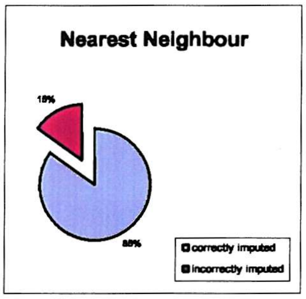




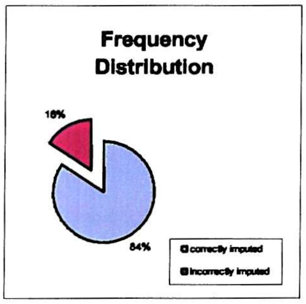
21
Terminal
Nodes

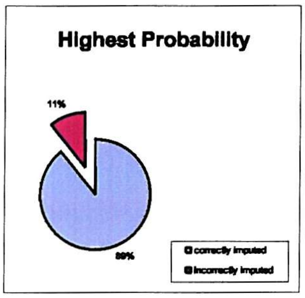


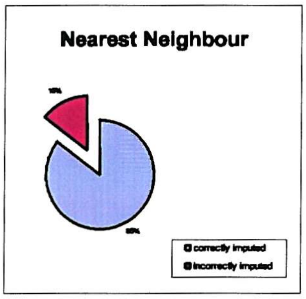




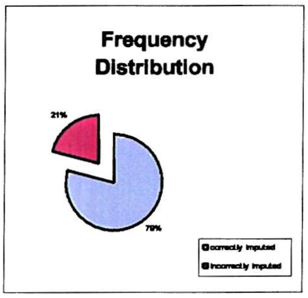
29
Terminal
Nodes

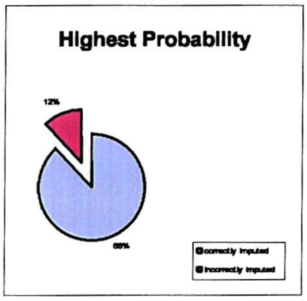


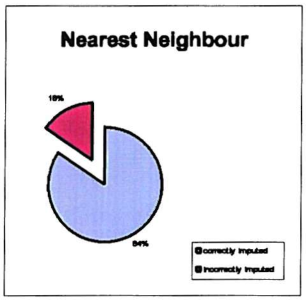




No
Tree



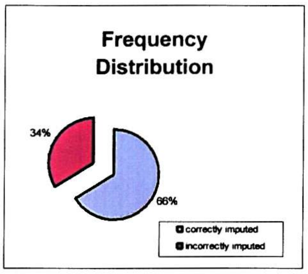


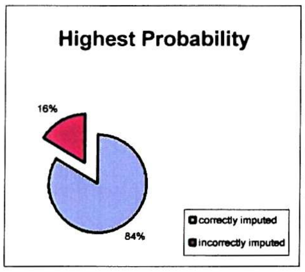


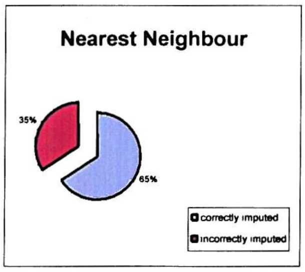
Appendix 6

Variable: Country of Birth - Ethnic
Accuracy of the imputation procedure

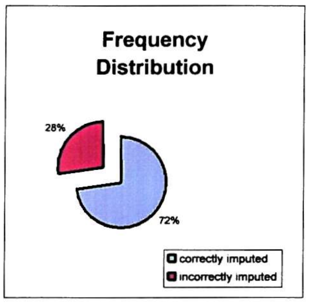
10 Terminal Nodes

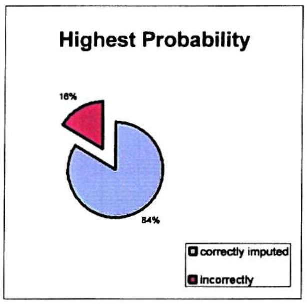


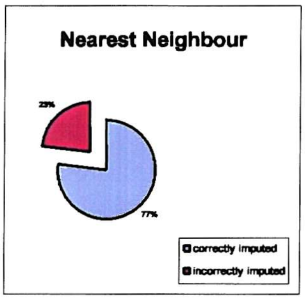




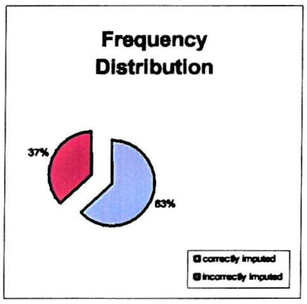
18 Terminal Nodes

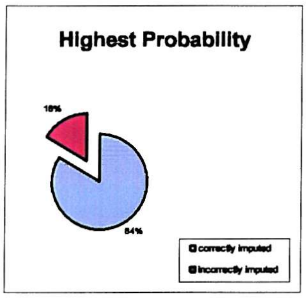


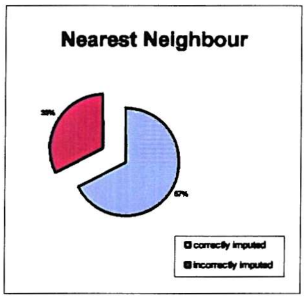




No Tree



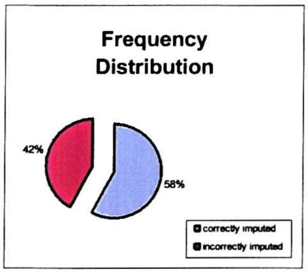


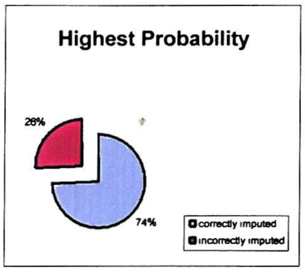


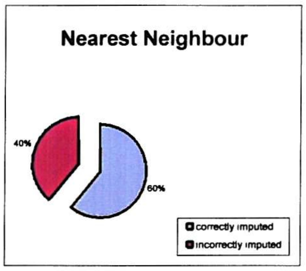
Appendix 6

Variable: Country of Birth - Long Term Illness
Accuracy of the imputation procedure

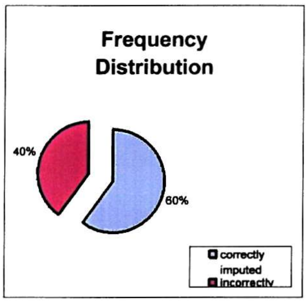
8
Terminal
Nodes

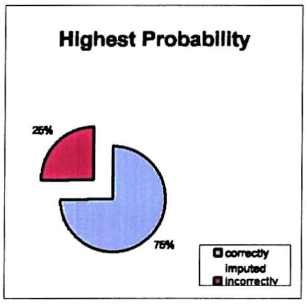


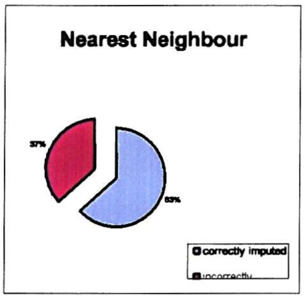




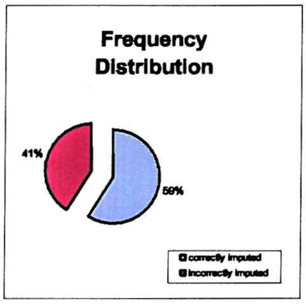
15
Terminal
Nodes

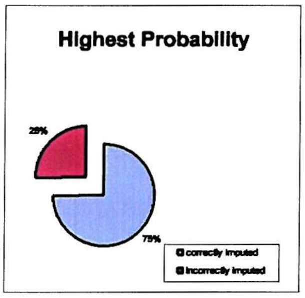


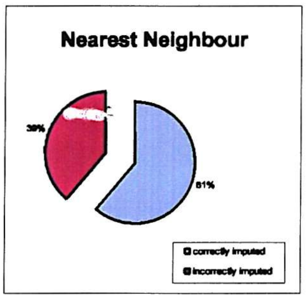




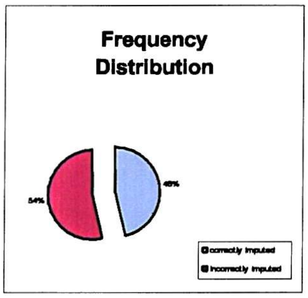
28
Terminal
Nodes

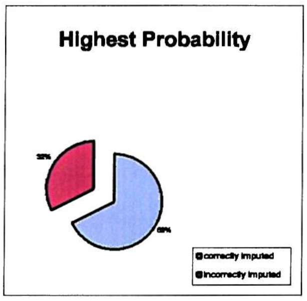


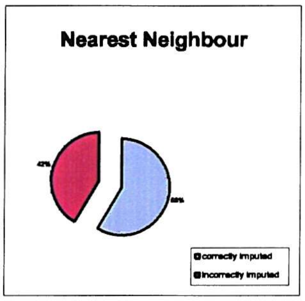




No
Tree





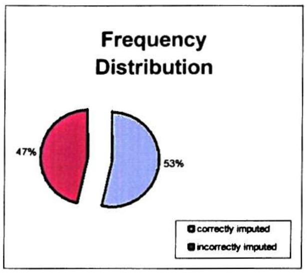


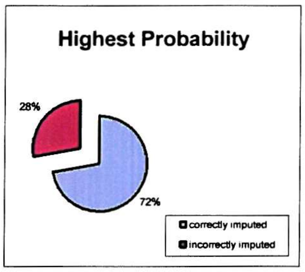
Appendix 6

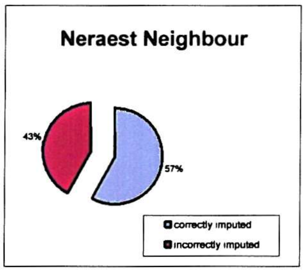
Variable: Ethnic - Long Term Illness

Accuracy of the imputation procedure

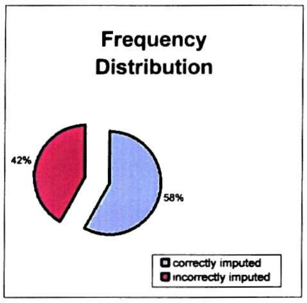
4 Terminal Nodes

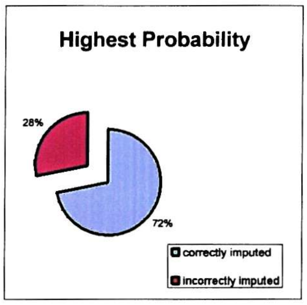


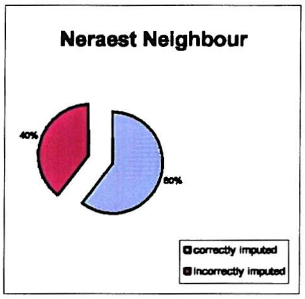




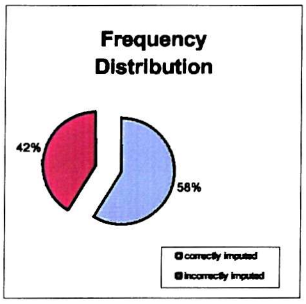
15 Terminal Nodes

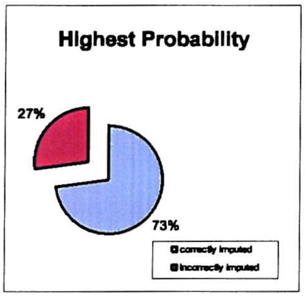


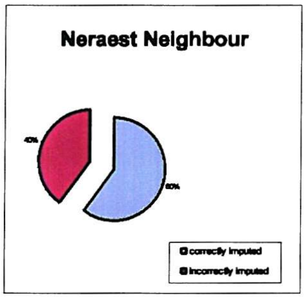




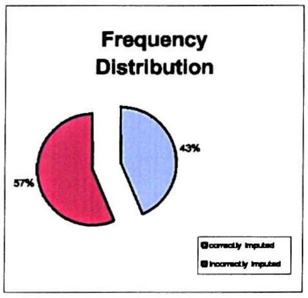
27 Terminal Nodes

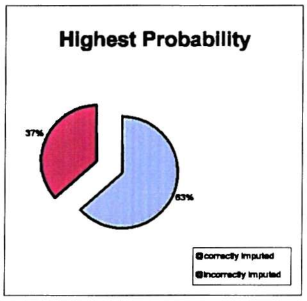


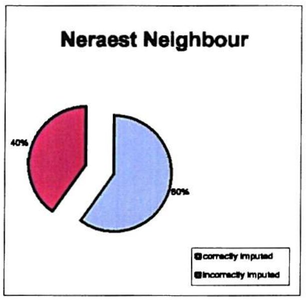




No Tree



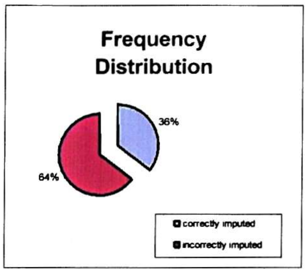


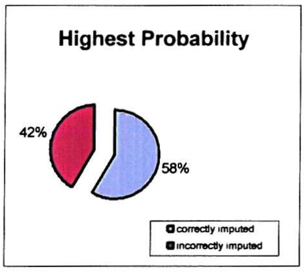


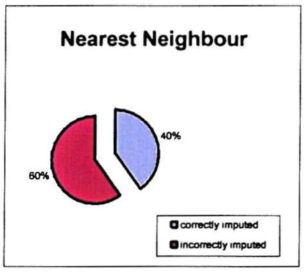
Appendix 6

Variable: Country of Birth - Ethnic - Long Term Illness
Accuracy of the imputation procedure

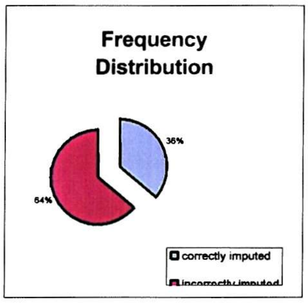
5 Terminal Nodes

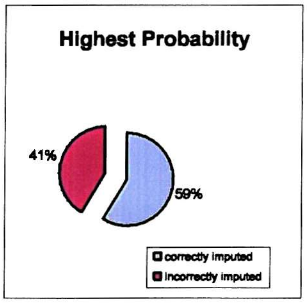


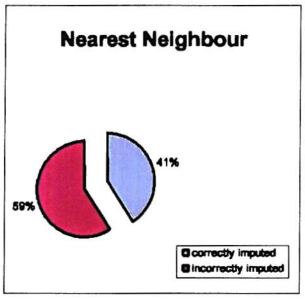




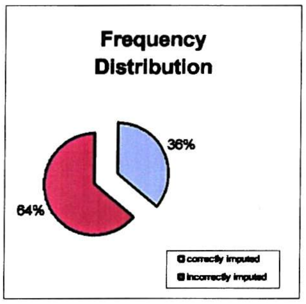
12 Terminal Nodes

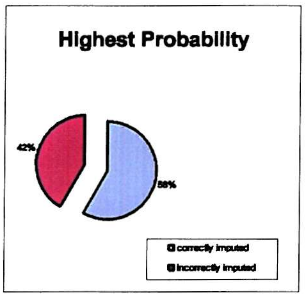


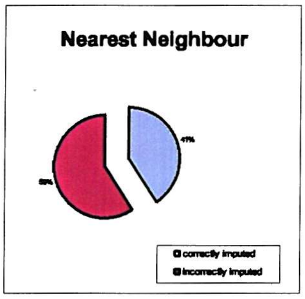




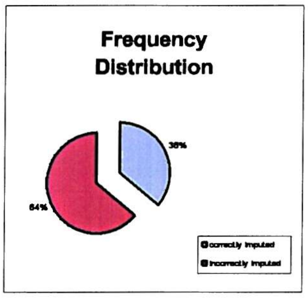
23 Terminal Nodes

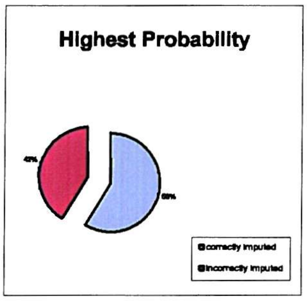


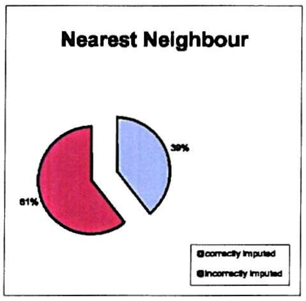




No Tree

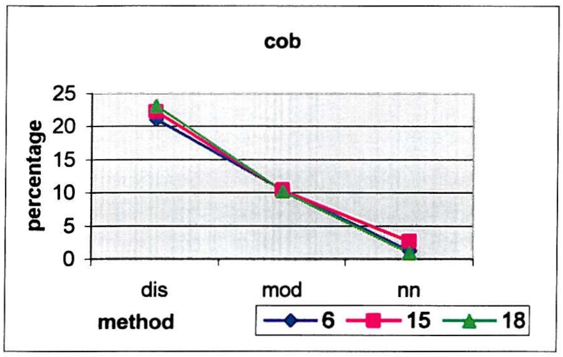


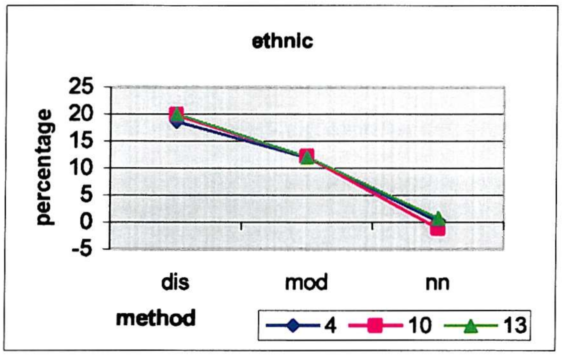


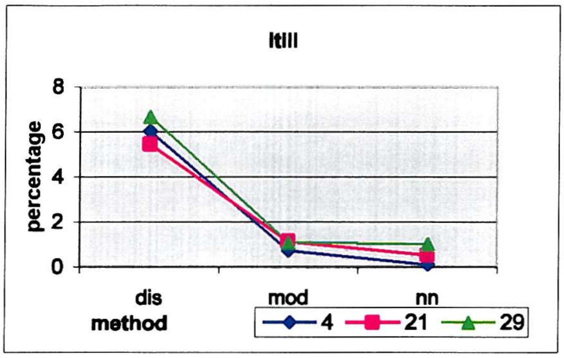


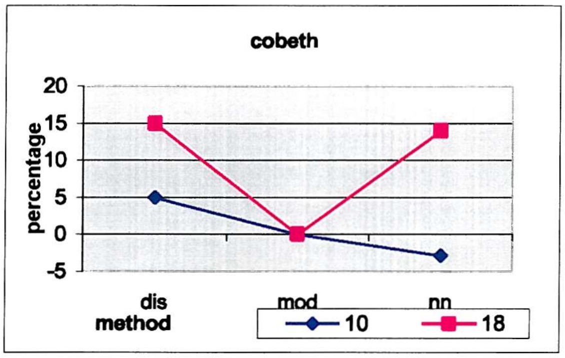
Appendix 7

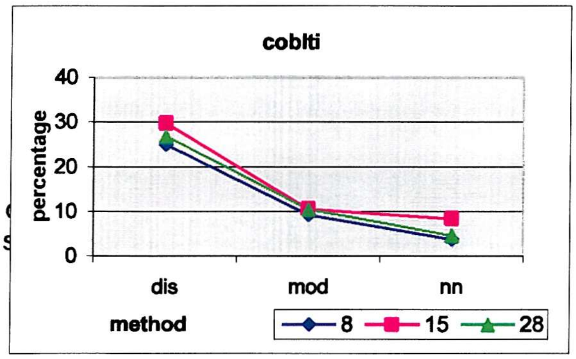
Improvement of the Percentage of Records Correctly Imputed with Respect to the Case of Not Using Trees (by variable, imputation method, and tree size)

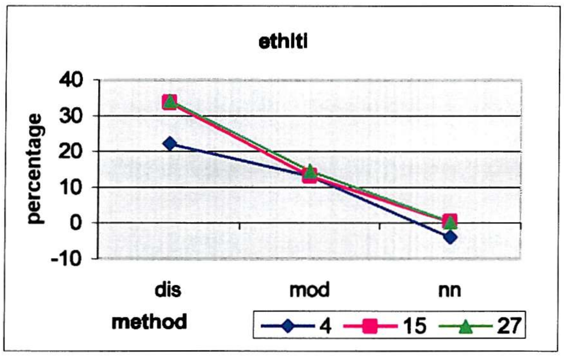


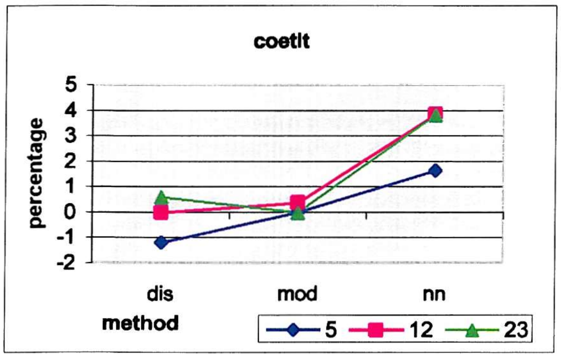






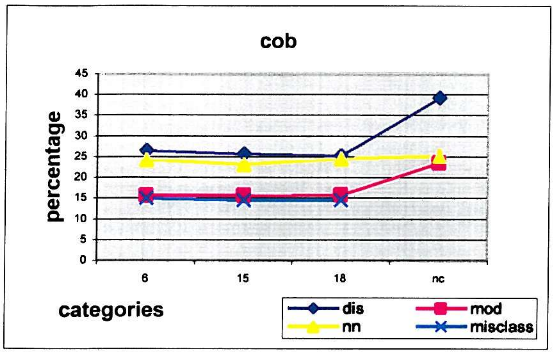


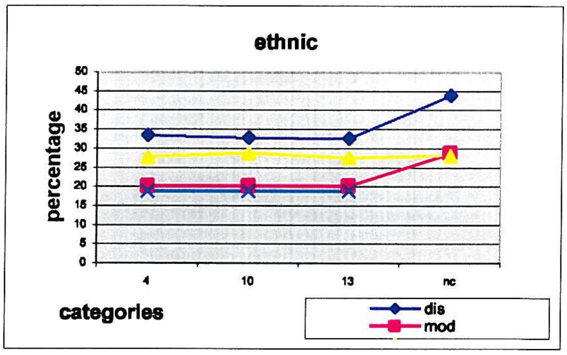


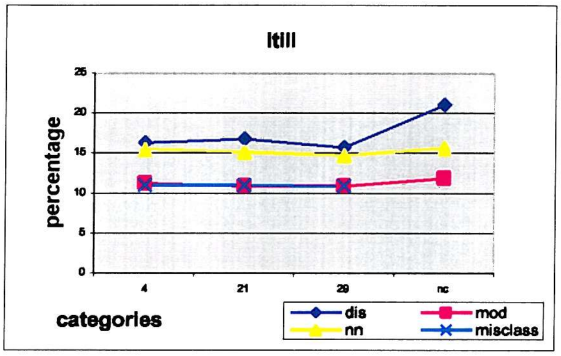


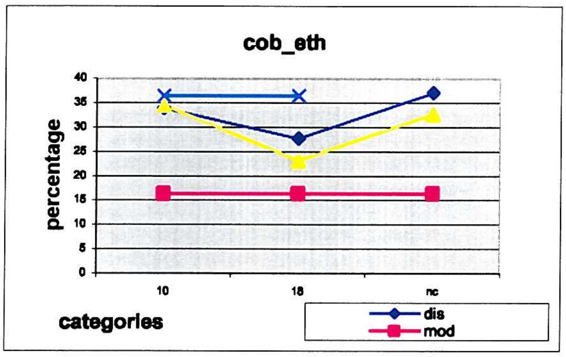
Appendix 8

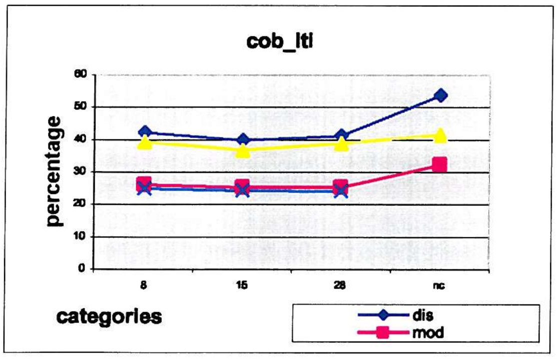
Relationship Between Misclassification Rates and Percentage of Records Incorrectly Imputed (by Variable, Tree Size and Imputation Methods)

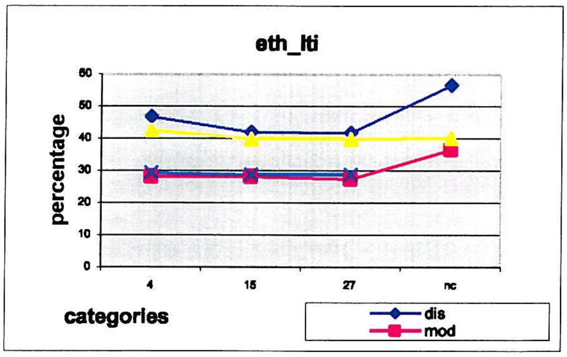


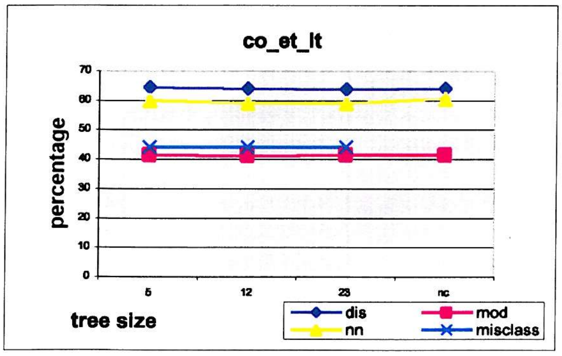












Appendix 9

Variable: Country of Birth

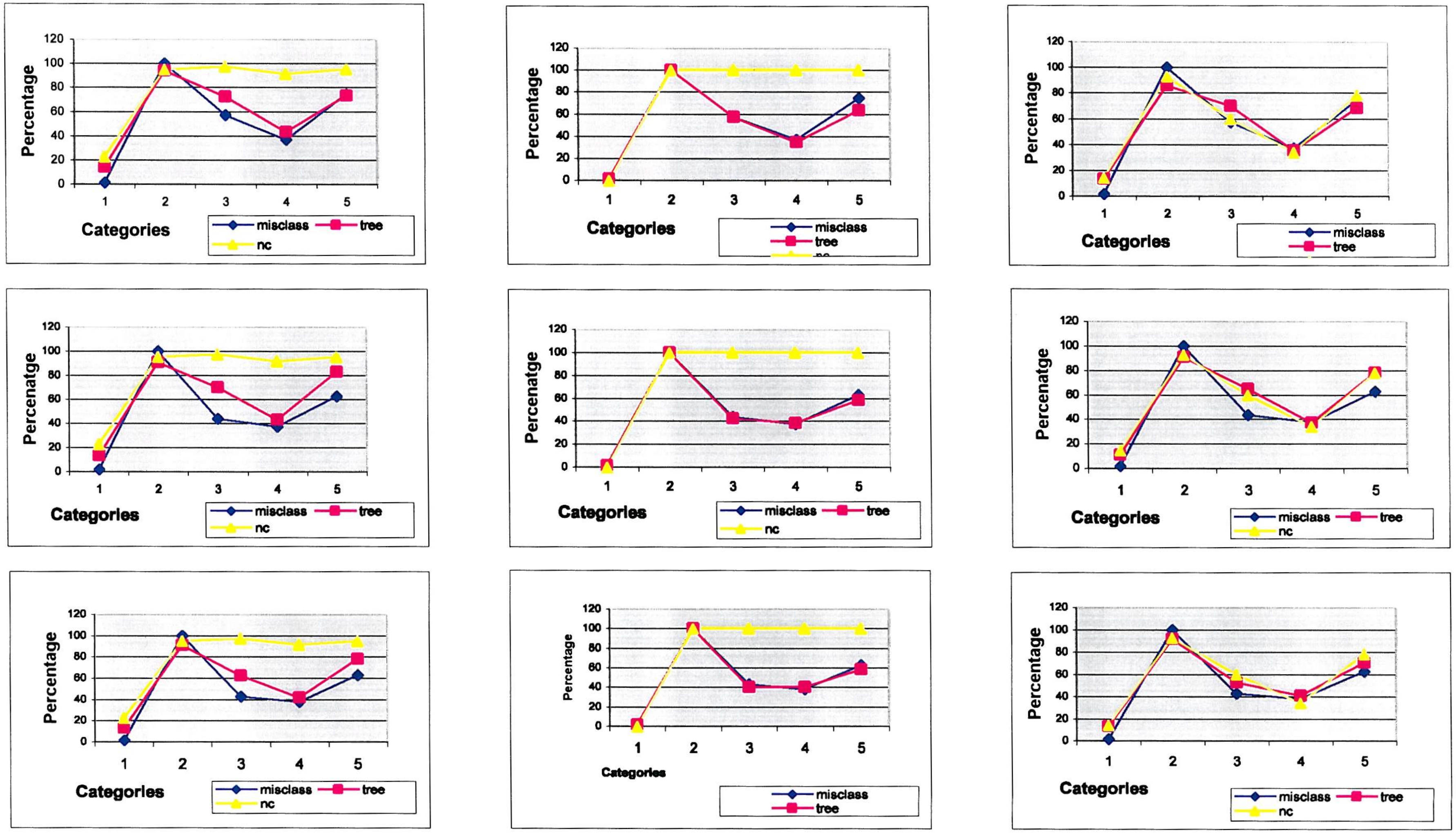
Misclassification rates against percentage of records incorrectly imputed (by tree size and imputation method)

FREQUENCY DISTRIBUTION — HIGHEST PROBABILITY — NEAREST NEIGHBOUR

6 Terminal Nodes

13 Terminal Nodes

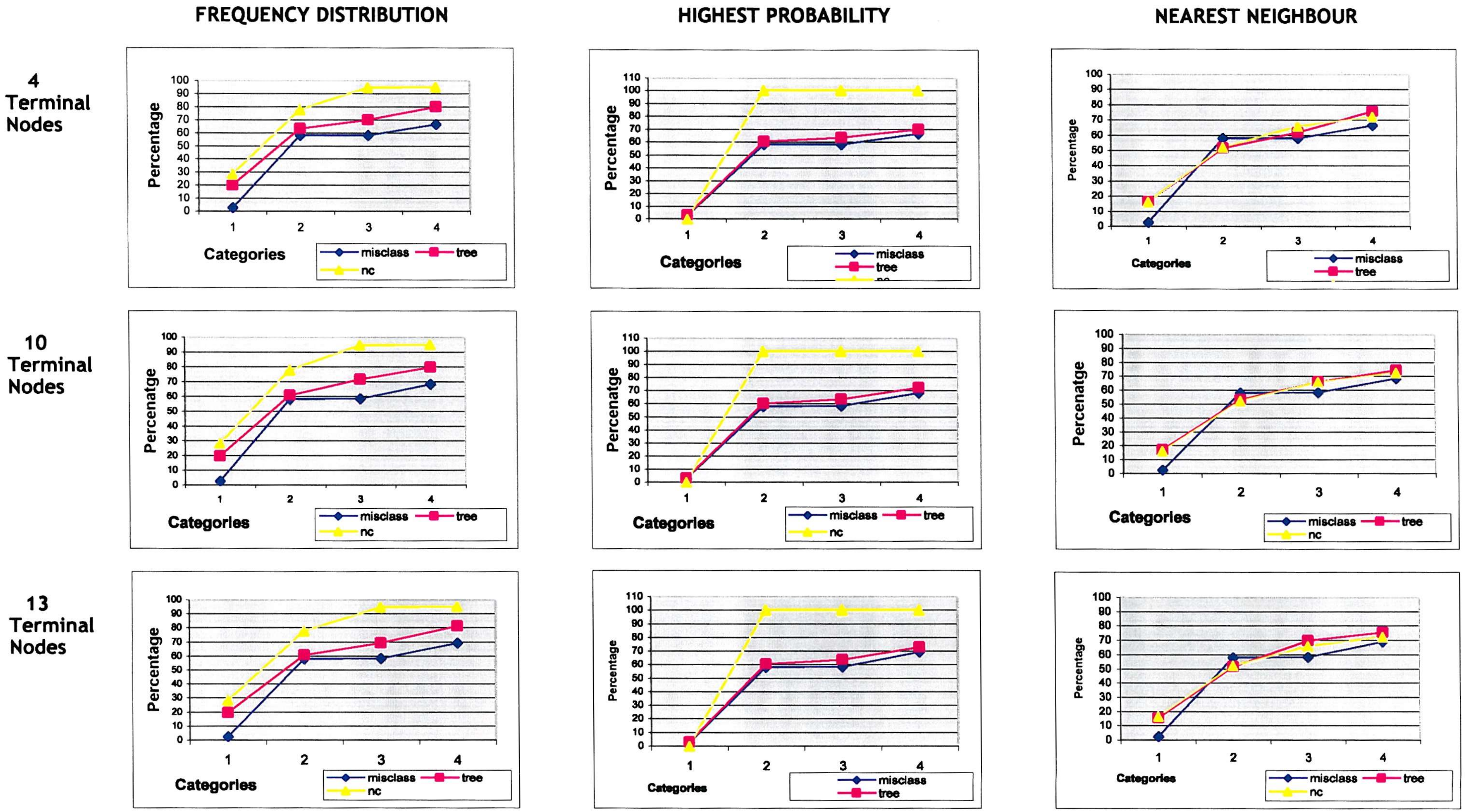
18 Terminal Nodes



Appendix 9

Variable: Ethnic

Misclassification rates against percentage of records incorrectly imputed (by tree size and imputation method)



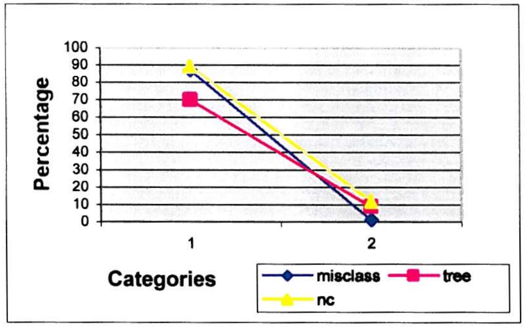
Appendix 9

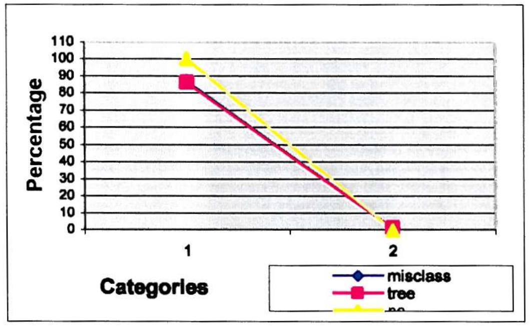
Variable: Long Term Illness

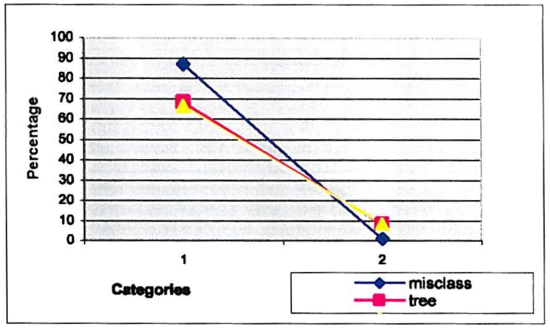
Misclassification rates against percentage of records incorrectly imputed (by tree size and imputation method)

FREQUENCY DISTRIBUTION | HIGHEST PROBABILITY | NEAREST NEIGHBOUR

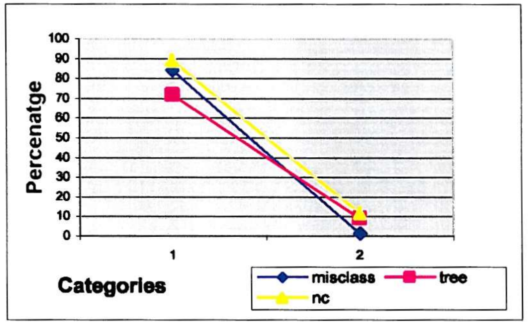
14 Terminal Nodes

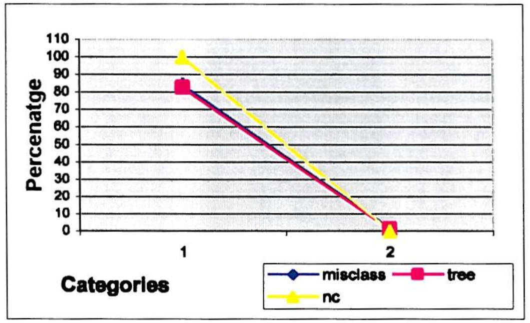


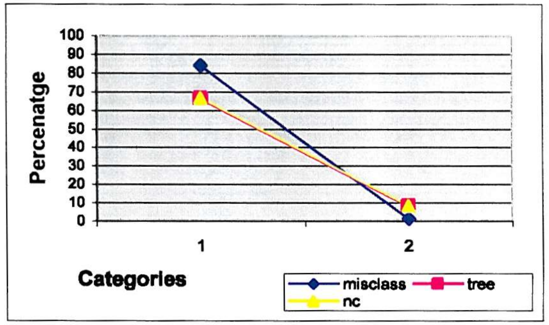




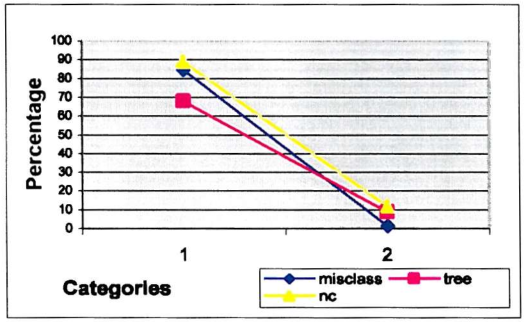
21 Terminal Nodes

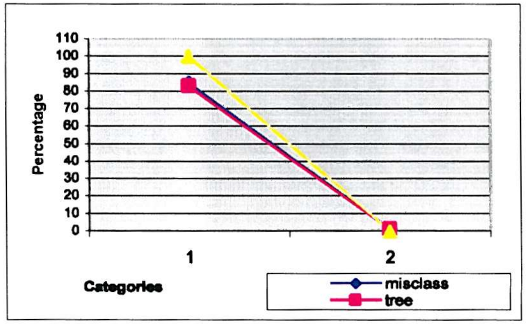


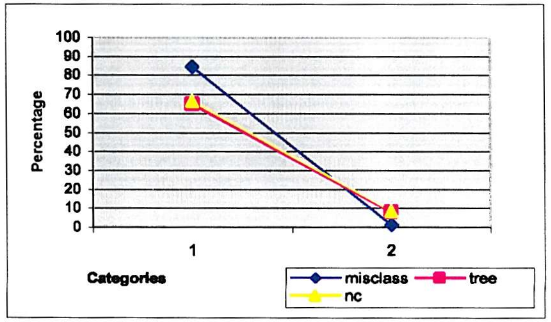




29 Terminal Nodes



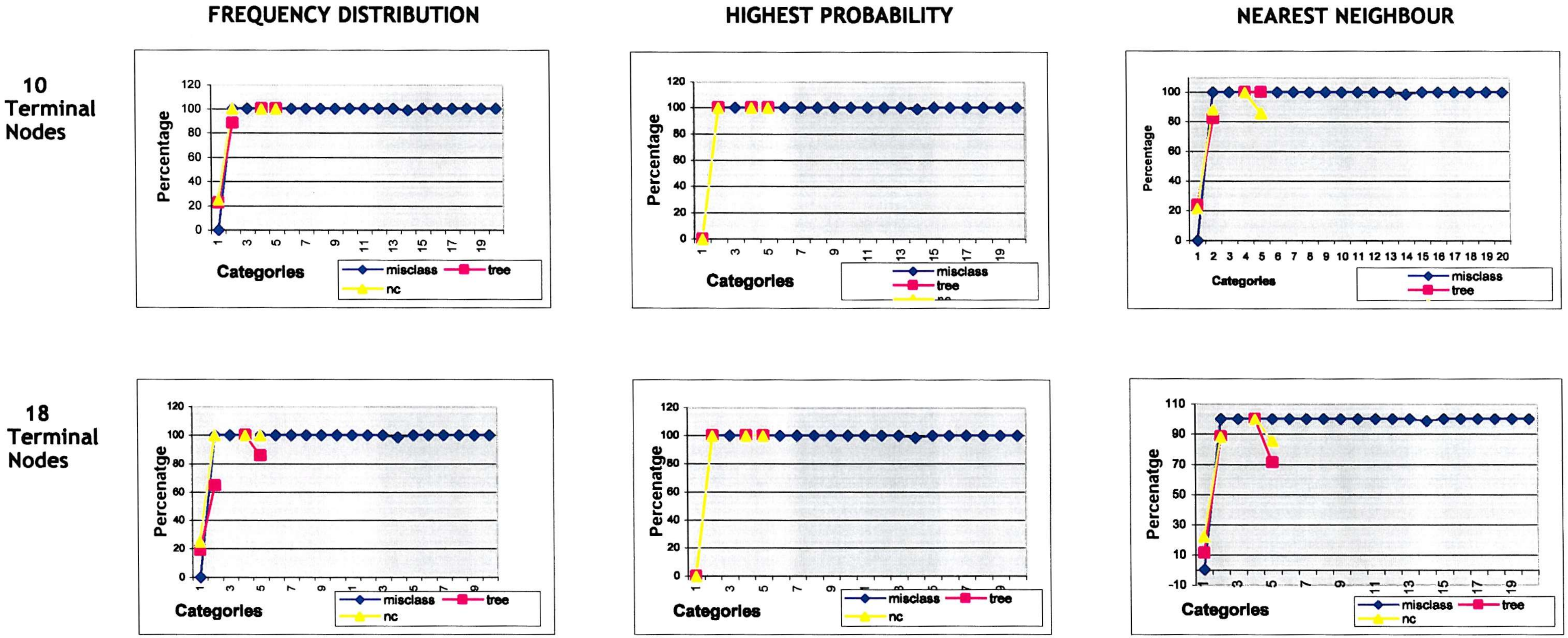




Appendix 9

Variable: Country of Birth - Ethnic

Misclassification rates against percentage of records incorrectly imputed (by tree size and imputation method)



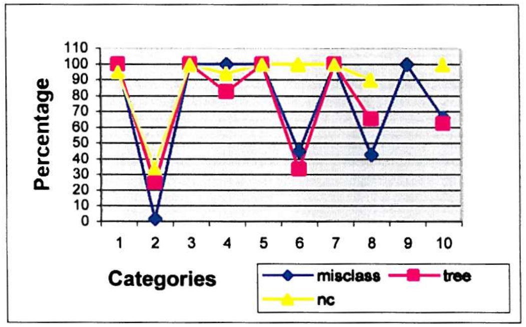
Appendix 9

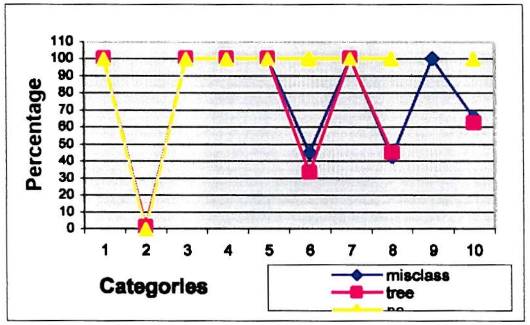
Variable: Country of Birth - Long Term Illness

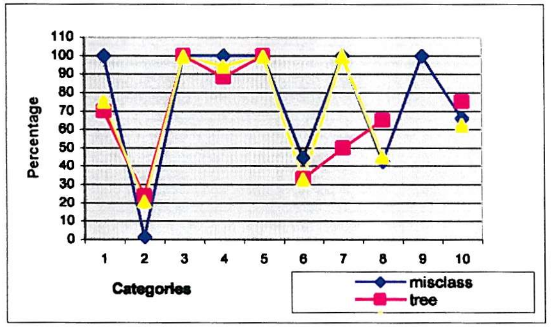
Misclassification rates against percentage of records incorrectly imputed (by tree size and imputation method)

FREQUENCY DISTRIBUTION | **HIGHEST PROBABILITY** | **NEAREST NEIGHBOUR**

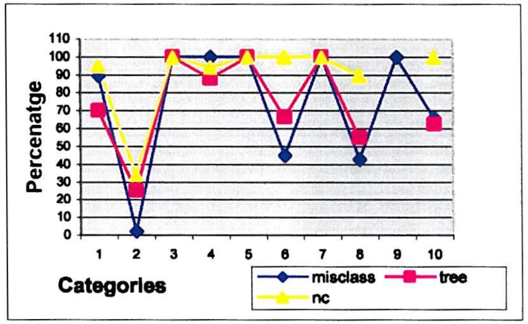
8 Terminal Nodes

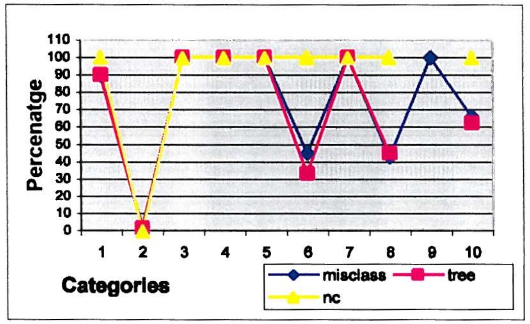


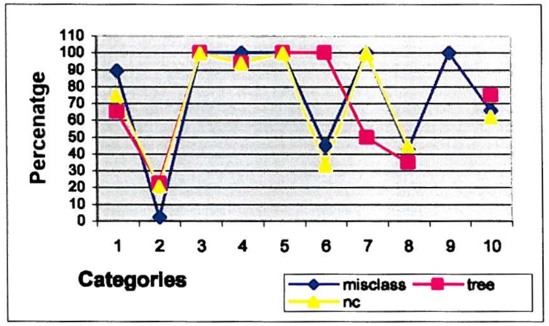




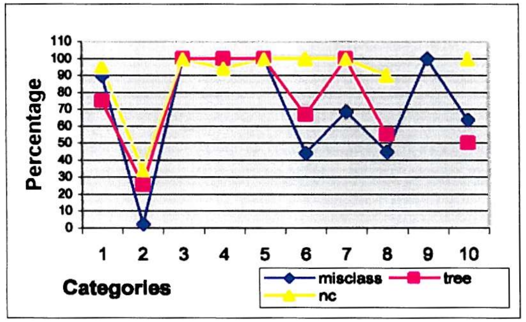
15 Terminal Nodes

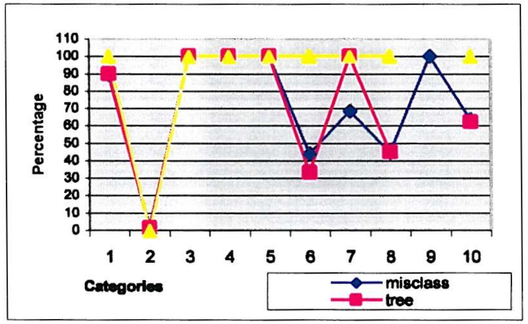


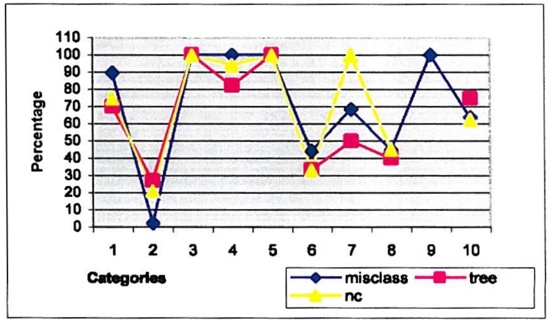




28 Terminal Nodes







Appendix 9

Variable: Ethnic - Long term Illness

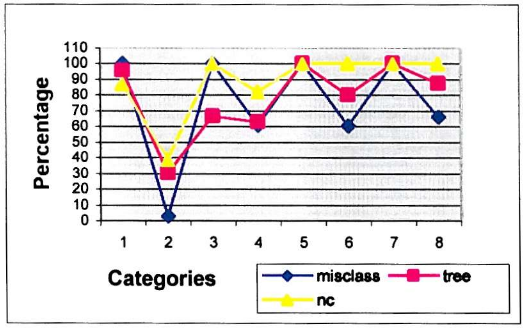
Misclassification rates against percentage of records incorrectly imputed (by tree size and imputation method)

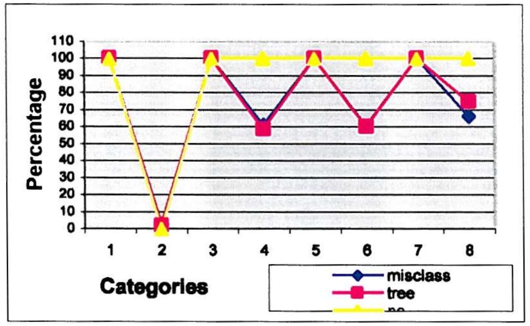
FREQUENCY DISTRIBUTION

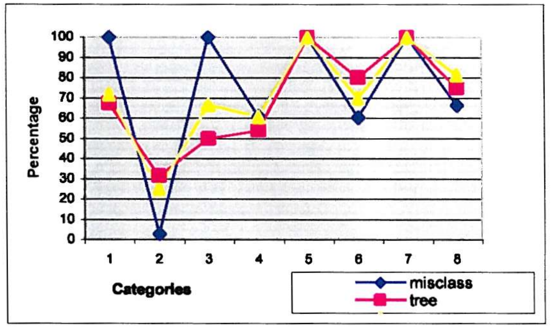
HIGHEST PROBABILITY

NEAREST NEIGHBOUR

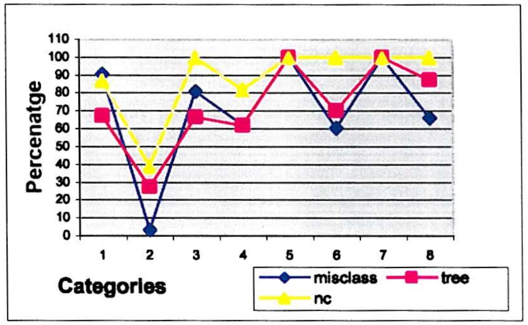
8 Terminal Nodes

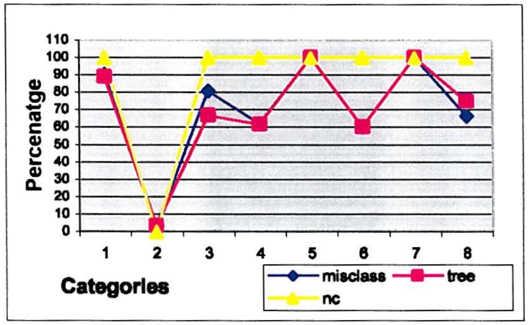


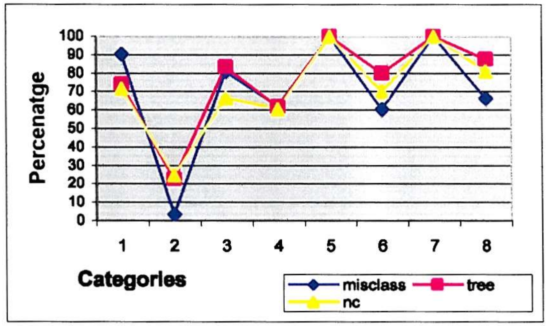




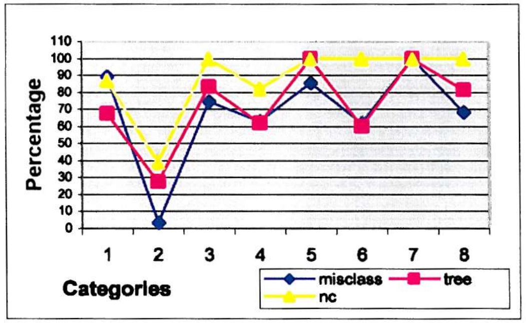
15 Terminal Nodes

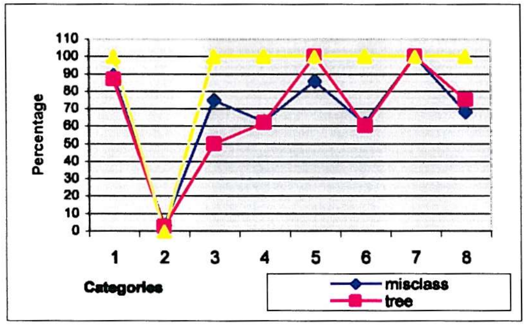


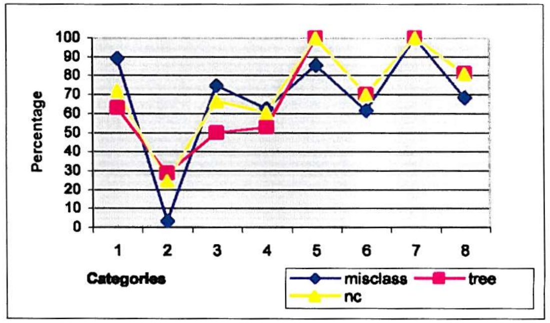




27 Terminal Nodes







REFERENCES

Anderson, F. & Whitfield, K. (2000) Editing and Imputation for the 2001 Census, *Internal Working Paper. Office for Nactional Statistics, UK.*

Bankier, M. (1999) Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extension for Future Censuses, *Working Paper submitted to the Conference of European Statisticians. Italy.*

Bankier, M., Fillion, J., Luc, M. & Nadeau, C. (1994) Imputing Numeric and Qualitative Variables Simulstaneously, *American Statistical Association. Proceedings of the Section on Survey Research Methods.*

Bankier, M., Houle, A.-M., Luc, M. & Newcombe, P. (1998) 1996 Canadian Census Demographic Variables Imputation, *American Statistical Association. Proceeding of Survey Research Methods.*

Bankier, M., Lachance, M. & Poirier, C. (1999) A generic Implementation of the New Imputation Methodology, *American Statistical Association. Proceeding of Survey Research Methods.*

Breiman, L. (1994) The 1991 Census ADjustment: Undercount or Bad Data?, *Statistical Science*, 9(4).

Breiman, L., Freidman, J.H., Olshen, R.A. & Stone, C.J. (1984) *Classification and Regression Trees* Pacific Grove, CA: Wadsworth.

Carson, R. (1984) Compensating for Missing Values and Invalid Responses in Contingent Valuation Surveys, *American Statistical Association. Proceeding of the Survey Research Methods.*

Chambers, R. (2000) Evaluation Criteria for Statistical Editing and Imputation, *Internal Working Paper, EUREDIT Project. http://www.cs.york.ac.uk/euredit/.*

Chen, J. & Shao, J. (1999) Jacknife Variance Estimation for Nearest Neighbour Imputation, *American Statistical Association. Proceeding of Survey Research Methods.*

Chen, J. & Shao, J. (2000) Nearest Neighbour Imputation for Survey Data, *Journal of Official Statistics*, 16(2).

Chen, J. & Shao, J. (2001) Jacknife Variance Estimation for Nearest Neighbour Imputation, *Journal of the American Statistical Association*, 96(453).

De Waal, T. (2000) New Developments in Automatic Edit and Imputation at Statistics Netherlands, *Statistical Commission and Economic Commission for Europe. Conference of European Statisticians. UN/ECE Work Session on Stastistical Data Editing.*

Ericksen, E., Kadane, J.B. & Tukey, J.W. (1989) Adjusting the 1980 Census of Population and Housing, *Journal of the American Statistical Association*, 84(408).

Fay, R.E. (1999) Theory and Application of teh Nearest Neighbour Imputation in Census 2000, *American Statistical Association. Proceeding of Survey Research Methods.*

Fellegi, I.P. & Holt, D. (1976) A Systematic Approach to Automatic Edit and Imputation, *Journal of the American Statistical Association. Proceeding of the Section on Survey Research Methods*, 71(353).

Gordon, A.D. (1987) A Review of hierarchical Classification, *Journal of the Royal Statistical Sociaty. Series A*, 150(2).

Granquist, L. (1997) The New View of Editing, *International Statistical Review*, 65(3).

Hand, D. (1997) *Construction and Assessment of Classification Rules* Wiley series in Probability and Statistics. John Wiley & Sons, Ltd.

Hill, C.J. (1976) A Report on the Application of the Systematic Method of Automatic Edit and Imputation to the 1976 Canadian Census, *Journal of the American Statistical Association. Proceedings of the Section on Survey Research Methods.*

Kalton, G. (1983) *Compensating for Missing Survey Data* Research Report Series. Survey Research Center. Institute of Social Research. The University of Michigan.

Kalton, G. & Kasprzyk, D. (1982) Imputing for Missing Survey Responses, *American Statistical Association. Proceeding of the Section on Survey Research Methods.*

Kalton, G. & Kasprzyk, D. (1986) The Treatment of Missing Survey Data, *Survey Methodology*, 12(1).

Kass, G.V. (1980) An Explanatory Technique for Investigating Large Quantities of Categorical Data, *Applied Statistics*, 29(2).

Kearney, A. & Ikeda, M. (1999) Handling of Missing Data in the Census 2000 Dress Rehearsal Integrated Coverage Measurement Sample, *American Statistical Association. Proceeding of Survey Research Methods.*

Lessler, J. & Kalsbeek, W. (1992) *Nonsampling Error in Surveys* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc.

Little, R.J.A. & D., R. (1987) *Statistical Analysis with Missing Data* Wiley Series in Probability and Mathematical Statistics.

Loh, W. & Vanischsetakul, N. (1988) *Journal of the American Statistical Association*, 83(403).

Madow, W.G., Olkin, I. & Rubin, D.B. (1983) (a) *Incomplete Data in Sample Surveys. Vol 1. Report and Case Studies* New York: Academic Press.

Madow, W.G., Olkin, I. & Rubin, D.B. (1983) (b) *Incomplete Data in Sample Surveys. Vol 2. Theory and Bibligraphies* New York: Academic Press.

Madow, W.G., Olkin, I. & Rubin, D.B. (1983) (c) *Incomplete Data in Sample Surveys. Vol 3. Proceedings of the Symposium* New York: Academic Press.

Martin, R.D. & Minardi, J. (1995) S-PLUS: Tools for Engineers and Scientists, *StatSci, A Division of MathSoft, Inc., Seattle, Washington.*

Mesa, D.M., Tsai, P. & Chambers, R. (2000) Using Tree-Based Models for Missing Data Imputation: An Evaluation Using UK Census Data, *Technical Report. AUTIMP Project.*

Poirier, C. (1999) A Functional Evaluation of Edit and Imputation Tools, *Statistical Commission and Economic Commission for Europe. Conference of European Statisticians. UN/ECE Work Session on Stastistical Data Editing.*

Poirier, C. (1999) A Comparison of Edit and Imputation Systems, *American Statistical Association. Proceeding of Survey Research Methods.*

Rancourt, E. (1999) Estimation with Nearest Neighbour Imputtaion at Statistics Canada, *American Statistical Association. Proceeding of Survey Research Methods.*

Rao, J.N.K. (2001) Variance Estimation in the Presence of Imputation for Missing Data.

Richards, J. (1999) DEIS System Requirements for Office for National Statistics, *Internal Report.*

Ryu, J.-B., Kim, Y.-W., Park, J.-W. & Lee, J.-w. (2001) Imputation Methods for the Population and Housing Census 2000 in Korea, *Bulletin of International Statistical Institute. 53rd Session Contributed Papers. TomeLIX, Book 2.*

Sande, I.G. (1982) Imputation in Surveys: Coping With Reality, *The American Statistician,* 36(3,1).

Schafer, J.L. (2001) The Practice of Multiple Imputation, *Docummemt Prepared for the Centre for Applied Social Surveys Short Course, Southampton.*

Series, U.K.G.S. (1996) *Report of the Task Force on Imputation* GSS Methodology Series No.3.

Silva, P.L. (2001) *Personal Communication.*

Sonquist, J.N., Baker, E.L. & Morgan, J.A. (1971) *Searching for Structure* Institute for Social Research. University of Michigan.

SPSS (1998) *AnswerTree 2.0 User's Guide* Chicago, IL: SPSS Inc.

Steel, P. & Fay, R.E. (1995) Variance Estimation fro Finite Populations with Imputed Data, *American Statistical Association. Proceeding of Survey Research Methods.*

Todaro, T.A. (1999) Overview and Evaluation of the AGGIES Automated Edit and Imputation System, *Statistical Commission and Economic Commission for Europe. Conference of European Statisticians. UN/ECE Work Session on Stastistical Data Editing.*

Valente, P. & Massimini, G. (199?) Processing the Italian Population and Housing Census Data., *Working Paper. Population and Housing Census and Territory Statistics Service. Italy.*

Vickers, P. (1999) DEIS System Requirements for Office for National Statistics, *Internal Report.*

Vickers, P. & Mohammed, Y. (1998) The Development and Evaluation of tye Donor Imputation System (DIS) for the 2001 UK Census of the Population and Housing, *Proceedings of the Joint IASS/IAOS Conference. Statistics for Economic and Social Development.*