

# **Imputation of Māori Descent for Electoral Calculations**

**Ian Westbrooke and Lisa Jones  
Analytical Support Division  
Statistics New Zealand**

Published in October 2000 by  
Statistics New Zealand  
Te Tari Tatau  
Wellington, New Zealand

---

Catalogue Number 01.095.0000  
ISBN 0-478-20758-1

# Contents

---

	Page
Summary	6
Why impute	7
What to impute	8
How to impute	9
Predictor variable selection for Maori descent imputation	9
Predictor variables selected for imputation	11
Variables not selected for imputation	13
What groups to use	15
Imputation	16
Evaluation	17
Conclusion	19
References	20

## List of tables

1 Household composition	11
2 Iwi	11
3 Ethnicity	12
4 Age group	12
5 Island	13
6 Māori descent from 1996 Census	17
7 Māori descent not allocated	18
8 Māori descent population with allocation - used for 1997 electoral calculations	18

## List of figures

1 Decision tree from CHAID analysis	10
-------------------------------------	----

---

**Blank Page 4**

# Acknowledgements

---

We wish to acknowledge the assistance of colleagues at Statistics New Zealand in developing this imputation and for editorial assistance. We would also like to thank Associate Professor Stephen Haslett, Massey University, for his review of our methodology and Whetu Wereta for her comments on the final draft of this paper.

## Summary

---

The Government Statistician decided that, for electoral purposes, Statistics New Zealand should impute Māori descent status for individuals not responding *Yes* or *No* to the Māori descent question in the 1996 Census of Population and Dwellings. *Imputation* is the allocation of a response based on the responses of others with similar attributes. Imputation provides a sounder basis for electoral population calculations than the approach used in 1994, when all who did not specify a clear *Yes* or *No* answer in the 1991 Census were effectively allocated to not being of Māori descent.

For the purposes of imputation, the key variables related to the Māori descent variable were identified using a statistical technique called CHAID (Chi-squared Automatic Interaction Detector). The imputation took place within subgroups created by cross-classification on the categories of the five variables identified – island, iwi, Māori ethnic group, Māori descent composition of the rest of the household, and age group. Within each subgroup, the proportions of those who specified a *Yes* or *No* response for Māori descent were used to allocate the remainder to *Yes* or *No*.

The effect of the imputation was to increase the proportion of those allocated to Māori descent from 16.0 percent to 17.4 percent of the total population. However, those imputed were allocated to Māori descent in a lower proportion than the proportion with Māori descent within the population who did specify *Yes* or *No*.

## Why impute

---

The Government Statistician is obliged under the Electoral Act 1993 to calculate the two electoral populations, General and Māori, based on the census ordinarily resident total and Māori descent population counts, and on the state of the electoral rolls at the end of the Māori Electoral Option (Māori Option) held after the census. These populations are used to calculate the number of electoral districts allocated to the Māori electoral population and the North Island General electoral population. The number of South Island General Electoral districts is fixed at 16. The electoral populations are also used to determine new electoral boundaries for all New Zealand electorates. In 1997 the Māori Option ran from 28 April to 27 August 1997, and the electoral populations were released in September 1997.

The quality and transparency of Statistics New Zealand's methodology for electoral population calculations is of critical importance. There is significant and widespread public interest in the electoral boundary revision process, especially in the calculation of the number of electoral districts. In 1994 the Māori Option was subject to legal challenge by Māori organisations, going all the way to the Privy Council.

In 1994, the Māori descent population was based only on those who answered *Yes* to the question on Māori ancestry in the 1991 Census (numbering 511,278). Significant numbers, however, answered *Don't know* or failed to specify an answer to this question, and these numbered over 252,000 in the 1991 Census.

After seeking and receiving legal advice, the Government Statistician decided that those not answering *Yes* or *No* to the question on Māori descent in the 1996 Census should be allocated to one of these two categories for electoral purposes. Allocation of responses was not used for the 1996 Census output, which maintains five categories: the three options from the census form *Yes*, *No*, *Don't know*; plus *Not specified* and *Unidentifiable*.

The special imputation creates a more accurate Māori descent population at the geographic levels required for electoral purposes. At the level of the North and South Islands, the Māori descent populations determine the number of electoral districts allocated to each electoral population. At electoral district level, the Māori descent populations are used by the Representation Commission to determine boundaries (Westbrooke and Ryan, 2000).

An imputation methodology was developed for assigning each person who did not answer *Yes* or *No* to the 1996 Māori descent question to one of these categories. The methodology needed to be soundly based, and able to be implemented readily in a short time. The imputation was performed on provisional 1996 Census results before the end of the 1996 calendar year. This allowed accurate estimates of the number of Māori electoral districts to be included in publicity material for the 1997 Māori Option. The methodology also needed to be adaptable for future census processing requirements, when the results would have to be calculated within a few months of census date to allow new electoral boundaries to be drawn.

## What to impute

---

The central output is a Māori descent variable which records *Yes* or *No* for each New Zealand resident recorded in the 1996 Census. The number without a definite answer to Māori descent, from this census, was over 355,000. These records needed to be imputed to one of the two categories of Māori descent, *Yes* or *No*.

We have imputed at the individual record level, as the following are needed for electoral calculations:

- Counts of Māori descent population for the North, South, and Chatham Islands, for the calculations that the Government Statistician carries out under the Electoral Act. These counts provide the basis for calculating the number of Māori and North Island General electoral districts, and the population quota for North Island General, South Island General, and Māori electoral districts. The outputs at this level are the most critical.
- Counts of Māori descent population for any proposed boundaries that the Representation Commission wishes to assess. (Each electoral district has a population of around 50,000.) Electoral districts can cut across all existing geographic boundaries, which is why we imputed, rather than estimating numbers of persons of Māori descent in geographical areas.

Counts down to meshblock level, each of which contains an average of about 100 individuals, are used in indicative calculations to assist at the beginning of the process. Electoral districts, existing or proposed, are the smallest units for which definitive counts are needed.

# How to impute

---

## Predictor variable selection for Māori descent imputation

A common method of imputation is conditional probability imputation. That is, given other information known about the respondent, what is the probability they will have a certain other characteristic? For example, if all we knew was a person's age, and that age is 1, we would assume about equal probabilities of their being male or female. However, if we knew their age to be 100, we would estimate the probability of their being female to be about 60%, and male 40%, based on 1996 Census figures.

The next step for the imputation is to identify good predictors of a respondent answering *Yes* or *No* to the question on Māori descent. That is, which other responses are available to use that correlate with a *Yes* or *No* answer to Māori descent. For those respondents who did not answer *Yes* or *No*, Māori descent status was then imputed from these predictors. The assumption is that the relationships that hold between the predictor variables and those who have responded *Yes* or *No* to Māori descent also hold for those who have not responded to either of these categories.

The bulk of the work involved analysing which are the best predictor variables of Māori descent (the dependent variable). As processing was not complete at the time of this work, a partial 1996 Census dataset containing over 1.6 million personal records, a little less than half the total, was used for variable identification.

Given the large number of possible predictor variables available in the Population Census, CHAID (Chi-squared Automatic Interaction Detector) analysis was used to choose the best subset of the variables. CHAID is a procedure for predicting the outcome of a dependent categorical variable, such as Māori descent, on the basis of a set of categorical predictor variables and is based on the chi-squared statistic. The method was first developed by Kass (1980).

CHAID requires all variables to be categorical including the dependent variable. Each variable can be either ordinal (discrete categories with an order, such as age: 27 is older than 26), or nominal (categories with no meaningful order such as ethnic origin).

The first step in CHAID analysis is to reduce the dimensions of the dataset. This is done by looking at each predictor variable, one at a time, to see if any of the categories of that variable can be grouped together. Each predictor variable is cross-tabulated with the dependent variable. The algorithm looks at each pair of the predictor variable categories and tests whether their behaviour is significantly different with respect to the dependent variable. If the two categories are not significantly different with respect to the dependent variable then the categories are merged into one. The significance test used is a chi-squared test of independence.

This new merged category is then tested against the remaining categories of that variable. An already merged category can be merged further with another category, and so on, until the most parsimonious configuration of that variable is found. This is done for each predictor variable in turn.

Once all the predictor variables have been optimally merged, the variable that is most significantly associated with the dependent variable is chosen as the best predictor. The dataset is then partitioned into subsets according to the chosen predictor variable categories. Each subset is then considered for further partitioning using the same algorithm that was applied to the entire dataset (Hawkins & Kass, 1982).

The process is repeated for each subset until some stopping criterion is met. CHAID was implemented using a SAS macro called Treedisc which can be obtained from the SAS Institute web site: <http://www.sas.com>. The main stopping criterion used by Treedisc is the significance level of the chi-squared test.

This recursive partitioning forms a tree structure. The ‘root’ of the tree is the entire dataset. The subsets and sub-subsets form the ‘branches’ of the tree. Subsets that are not partitioned any further are the ‘leaves’. This decision tree can be requested as a SAS output.

Figure 1

### Decision tree from CHAID analysis

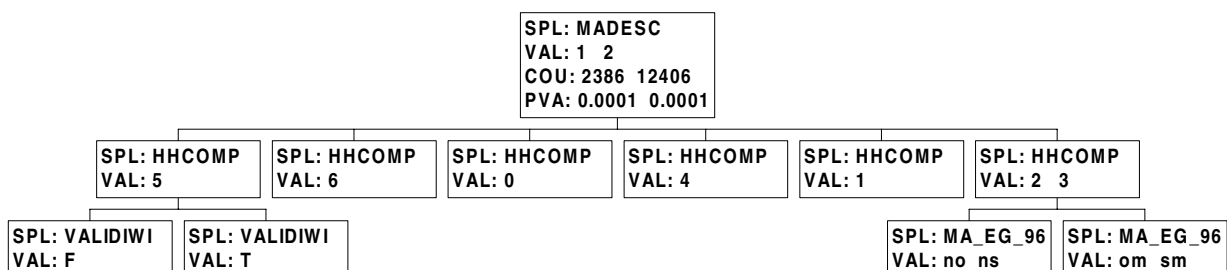


Figure 1 is an example of part of a decision tree produced using CHAID analysis on a sample of the census data that had specified *Yes* or *No* responses to the Māori descent question. The same analysis was run over several different samples for cross-validation purposes.

The top box contains four items:

- the variable on which the split is occurring (SPL:MADESC), that is, Māori descent
- the categories of the variable (VAL 1=*Yes*; 2=*No*.)
- the count of that subdivision that goes to each category of the dependent variable (COU: 2386 answered *Yes* to Māori descent, 12406 answered *No* to Māori descent)
- the p-value (significance level) of the chi-squared test from the split (PVA).

This example has been simplified by removing some of the splits and information from each box, excluding the top box.

The household composition variable (HHCOMP), which is explained later, is the first variable on which a split occurs, indicating that it is the best predictor of Māori descent. Note that the responses 2 and 3 for household composition have been combined by CHAID using the first step of the CHAID process as outlined above.

## Predictor variables selected for imputation

### *Household composition (HHCOMP)*

It was important to exploit information available from natural groupings of individual records when there was more than one person in the household. As the large majority of households in New Zealand consist of people that are related, it was expected that a good predictor of whether a person was likely to answer *Yes* or *No* to the question on Māori descent would be responses of others in the household to the same question. This variable summarises Māori descent information from the entire household. Seven separate categories were used to summarise the available information, as shown in Table 1.

Table 1

### Household composition

HHCOMP	Description
0	No information due to: household size of 1 non-private household no information on Māori descent for others in household
1	Household size of 2 with other member being of Māori descent
2	Household size of 3 or more with all others members being of Māori descent
3	Household size of 3 or more with at least half of other members being of Māori descent
4	Household size of 3 or more with less than half of other members being of Māori descent
5	Household size of 3 or more with no others having Māori descent
6	Household size of 2 with other member not being of Māori descent

This variable came out as being the strongest indicator of Māori descent in the CHAID analysis. The information will be available early in future census processing which is vital if the imputation methods are going to be consistent across time.

### *Iwi (VALIDIWI)*

In the final census data, an individual only has an iwi recorded if they have replied *Yes* to Māori descent. However, the provisional file showed all responses coded for iwi. Many of those needing to be imputed had recorded an iwi such as 'Ngapuhi'. As expected, having a valid iwi code provided a good indicator of Māori descent status. The valid iwi variable for the imputation, derived especially from census files prior to final editing, had two categories, as shown in Table 2.

Table 2

### Iwi

VALIDIWI	Description
True	Valid iwi present
False	Valid iwi not present

### **Ethnicity (MA\_EG\_96)**

Ethnicity proved through CHAID analysis to be a very good indicator of Māori descent status also. Ethnicity was originally split into four categories:

- no Māori ethnic group (*no*)
- no ethnic group specified (*ns*)
- specified Māori ethnic group plus other ethnic group/s (*om*)
- only specified Māori ethnic group (*sm*).

The CHAID analysis almost always grouped the 'no Māori' category with 'not specified' and also grouped the two Māori categories together. Therefore the ethnic group variable followed these groupings and collapsed the variable into two categories as shown in Table 3.

Table 3

### **Ethnicity**

<b>MA_EG_96</b>	<b>Description</b>
Yes	No Māori ethnic group or ethnic group not specified.
No	Specified Māori ethnic group only or specified Māori ethnic group plus other ethnic group

### **Age group (AGEGP)**

There is an obvious age structure within Māori descent status. The population of Māori descent has a lower median age and is more skewed to younger ages than the non-Māori descent population. Because consecutive ages were similar, and in order to reduce the number of categories, the ages were best handled in groups rather than leaving each individual age as a category.

Age grouped by 10 years, as shown in Table 4, proved to be a good discriminator of Māori descent status when included in the CHAID analysis.

Table 4

### **Age group**

<b>AGEGP</b>	<b>Description</b>
0–9	From 0 to 9 years of age
10–19	From 10 to 19 years of age
20–29	From 20 to 29 years of age
30–39	From 30 to 39 years of age
40–49	From 40 to 49 years of age
50–59	From 50 to 59 years of age
60–69	From 60 to 69 years of age
70+	70 years of age or over

## **Island (ISLAND)**

Because of the critical nature of the Māori descent population count at an island level, all imputation was done separately for the North and South Islands. Chatham Islands was included with the North Island because it has a high proportion of Māori in the population, making it similar to the North Island.

Table 5

### **Island**

<b>ISLAND</b>	<b>Description</b>
North	North Island or Chatham Islands
South	South Island

On the basis of the CHAID analysis the final variable selection was:

- household composition (7 categories)
- iwi (2)
- Māori ethnicity (2)
- age group (8)
- island (2)

giving 448 possible subgroups.

## **Variables not selected for imputation**

In addition to the variables selected above, many other census variables were investigated but did not prove to be good indicators of Māori descent status for various reasons:

- language (presence of Māori language or not)
- religion (presence of Māori religion or not)
- birthplace (New Zealand or not)
- sex
- total income
- labour force status
- regional council areas
- urban/rural.

The first three in the list did show up in the lower levels of the CHAID analysis, but generally there were not enough people who had a particular response for that variable to make a significant difference if they were included. The others barely showed up, if at all. It is possible that if the variables identified in the previous section as good predictors were not available, some of these other variables would be identified as good predictors. However they are not useful here as they are highly correlated to the selected variables, so most of their predictive power has been eliminated with the selection of the correlated variables.

Family records were also examined to see how much extra information could be obtained from other family members. For example, if either parent indicated they were of Māori descent it is likely that any children would be of Māori descent also. By looking at each family coded, separate variables were created which contained information on whether each individual was a child of a Māori descent parent and whether they were a parent of a Māori descent child. As expected, this came out as significantly related to Māori descent, and could be as good a predictor compared with those chosen, notably household composition. Although family coding includes potentially important information it was not practical to include it in the imputation process for the following reasons:

- There may be very limited time between the 2001 Census and the deadline for electoral outputs. Family coding is complex and when the electoral calculations have to be made, the family information will not be available. Ensuring consistency in our imputation procedures rules out the use of family coding. Former Deputy Government Statistician and Representation Commission member Ron Welply stated in a review of electoral methodology papers in relation to Māori descent imputation: “Part of the (Government) Statistician’s process is to ensure not only that a practice is practicable now but remains so in the future.”
- The household variable performed well as a substitute. Living together in a household is less directly related to Māori descent than belonging to the same family, for example as a child of a person of Māori descent. However, the household information is readily available for the process, and is applicable to more individuals than family information. The time and resource constraints of this study precluded testing household versus family information.

## What groups to use

---

Having chosen the variables for the imputation, we had two choices for groups on which to base the imputation. One was to base the imputation on the subgroups determined by the CHAID algorithm. However, experience with results from CHAID on different samples of the data indicate that although the variables involved remained the same or very similar, the particular subgroups could vary distinctly. The alternative was to take the variables selected from the CHAID analysis, and create the Cartesian product of the categories selected for each of these variables; that is take all 448 combinations of variable values. We chose the second as it was conceptually simpler, more robust, and easier to explain. The only possible drawback was that some groups could have very small populations. If there are few with specified Māori descent, then imputation can become difficult. This did not appear to be a problem with the chosen variables, once the number of age groups was reduced to eight, ending with the 70 years plus category. Choosing the full Cartesian product with its large number of groups couldn't make the imputation any less accurate than choosing only a limited number of groups, and could improve it.

# Imputation

---

The full census data were summarised into each of the groups. The information on the allocation of those specified was used to allocate the remainder to either *Yes* or *No*.

There was a choice either to take each group and allocate all its unallocated members on a deterministic basis to either *Yes* or *No*, or to allocate them proportionately (using a random number process) to *Yes* or *No* according to the proportions in that group. Experience with a deterministic approach showed that it had potential for bias, so the proportional option was chosen.

In addition, there were almost 100,000 additional personal records with almost no information that had been included after the evaluation dataset was created. These records, called dummies, were created when there was strong evidence a person should be included in the census count, but where no census form existed. Typically, no information was available for allocating dummies, although some had information for the Māori descent of other people in the same household. For those with no information, a second stage was added to the allocation, assigning Māori descent randomly to these records according to the proportions in each island at the end of the first stage allocation.

## Evaluation

---

The variable selection approach consistently identified the same variables related to Māori descent over several samples. Age group was generally low in the hierarchy of variables selected. It was nevertheless included because the pattern of greater Māori descent at lower ages was consistent in most subgroups we examined. The inclusion of a North Island/South Island split was decided beforehand because of the key nature of the electoral outputs at this level, and because of the distinct difference in the Māori populations between the islands. Both island and lower geographic variables (16 regional council areas and an urban/rural split) showed up as less important than the variables chosen. Since the island split does seem to be important in the final results, this raised some concerns whether further geographic breakdown might add to the quality of the imputation. However, the chosen subgroups already numbered 448, and some of them were becoming very small.

The overall assessment is that a reasonable balance was achieved by including enough variables to provide a good imputation, but not creating too many categories, especially nearly empty ones.

The relative split of Māori descent within the subgroups shows that most of the subgroups were strongly polarised, with the vast majority of those specified being in one group, either *Yes* or *No*. For the 1.6 million census records available for evaluation, almost 105,000 needed allocation. Of these, 81 percent were in groups where more than 19 out of 20 of those specified gave *No* to Māori descent, and another 7 percent where more than 19 out of 20 specified *Yes*.

Almost all those needing allocation in the evaluation data had additional information clearly related to Māori descent in the form of one or more of: a valid iwi present, ethnic group information (Māori either clearly present or absent) or information on the Māori descent composition of the rest of the household. For the partial evaluation data, less than 3,000 (under 3 percent of those to be allocated), had no iwi, ethnicity not specified, and no household information. The results of the imputation on the final census data are shown in Tables 6–8. *Note:* Numbers in Tables 6, 7 and 8 have been rounded to the nearest 10 to preserve confidentiality. Due to rounding, the components do not always add exactly to the given totals.

Table 6

### Māori descent from 1996 Census

	Yes	No	Total allocated	Not allocated	Total
North Island	501,860	1,942,980	2,444,840	274,080	2,718,920
Yes/No split	20.5%	79.5%	100.0%		
South Island	77,860	740,590	818,450	80,940	899,380
Yes/No split	9.5%	90.5%	100.0%		
Total	579,710	2,683,580	3,263,290	355,010	3,618,300
Yes/No split	17.8%	82.2%	100.0%		

Table 7

**Māori descent not allocated**

	Stage 1 allocation		Stage 2 allocation		Total allocation	
	Yes	No	Yes	No	Yes	No
North Island	28,230	171,880	14,930	59,040	43,150	230,920
Yes/No split	14.1%	85.9%	20.2%	79.8%	15.7%	84.3%
South Island	4,240	62,120	1,320	13,260	5,560	75,380
Yes/No split	6.4%	93.6%	9.1%	90.9%	6.9%	93.1%
Total	32,460	234,000	16,250	72,300	48,720	306,300
Yes/No split	12.2%	87.8%	18.4%	81.6%	13.7%	86.3%

Table 8

**Māori descent population with allocation  
- used for 1997 electoral calculations**

	Yes	No	Total
North Island	545,010	2,173,910	2,718,920
Yes/No split	20.0%	80.0%	100.0%
South Island	83,420	815,970	899,380
Yes/No split	9.3%	90.7%	100.0%
Total	628,430	2,989,870	3,618,300
Yes/No split	17.4%	82.6%	100.0%

The additional information has led to the allocation of a significant proportion of respondents as having Māori descent, but not at the same rate as among those who were already specified. At the overall level, 17.8 percent of those who specified *Yes* or *No* had Māori descent, but only 13.7 percent were imputed to Māori descent. Those not of Māori descent may be more likely to give a *Don't know* response or fail to specify an answer to the Māori descent question than those of Māori descent.

## Conclusion

---

Imputation of Māori descent using carefully selected census variables has yielded Māori descent population counts which are more accurate than previous counts, thus providing a significantly better and fairer basis for electoral population calculations.

The effect of the imputation is to increase the proportion of those allocated to Māori descent from 16.0 percent to 17.4 percent of the total population.

## References

---

- Breiman L, Friedman J H, Olshen R A & Stone C H (1984), *Classification and Regression Trees*, Wadsworth: Belmont, CA.
- du Toit S H C, Steyn A G W & Stumpf R H (1986), *Graphical Exploratory Data Analysis*, Chapter 8, Springer-Verlag: New York.
- Hawkins D M & Kass G V (1982), Automatic Interaction Detection, in Hawkins D M, ed, *Topics in Applied Multivariate Analysis*, 267–302, Cambridge Univ Press: Cambridge.
- Kass G V (1980), An Exploratory Technique for Investigating Large Quantities of Categorical Data, *Applied Statistics*, 29, 119–127.
- Westbrooke, I & Ryan, MP (2000), *The Mathematics of Electoral District Allocation In New Zealand*, Statistics New Zealand Research Report #12.