Understanding and Working with Ethnicity Data

A technical paper

Statistics New Zealand

April 2005

Contents

Figures

Tables

Introduction

This technical paper provides support to users of ethnicity data. It addresses issues arising from the Review of the Measurement of Ethnicity and advises on best practices. This paper was prepared by Robert Didham with assistance from Deb Potter and Jo-anne Allan. Among the topics covered are: ethnic mobility, contextual effects and the dynamics of ethnicity; sources of data; collection issues including consistency, and reduction of multiple responses; comparing collections; output choices, time series analysis, and data integration. This paper does not offer any discussion on ethnicity-based funding models. It would be helpful to read this paper in conjunction with The Statistical Standard for Ethnicity and When Individual Responses Exceed Input Storage - A Procedure For Unbiased Reduction. Both are available on the Statistics New Zealand website at www.stats.govt.nz.

Background

In June 2004 Statistics New Zealand released a report the *Review of the Measurement of Ethnicity.* This report was the result of consultation with users and producers of ethnicity data, as well as, contributions from respondents and those working in the area of survey design and development. The review found that ethnicity as a concept remains a key social variable. Users made it clear that there was a need for the ongoing collection of detailed ethnicity data.

They agreed that ethnicity should be measured in a consistent way across all official statistics, in order that:

- all collections of official statistics measuring ethnicity should have the capacity to record multiple ethnicity responses

- the method of reporting ethnicity in all collections of official statistics be self-identification

- 'New Zealander', 'Kiwi' and like responses be separately classified

- the practice of prioritising ethnic group responses to one per individual be discontinued.

While most major suppliers of ethnicity data have moved away from the output of prioritised ethnicity data, some producers and users have continued to prioritise ethnicity data to one response per person where multiple responses have been provided. With continued increases in multiple responses, especially among younger respondents, the need to retain multiple responses in collections has become increasingly important.

Ethnic mobility, contextual effects and the dynamics of ethnicity

Ethnicity is not fixed. People in New Zealand, as in other countries, may change the ways in which they identify themselves over time or they may identify themselves differently in different environments. Many aspects of an individual's circumstances affect how they identify their ethnicities and this may differ markedly from how a third party might identify them. Some of these aspects are important for the interpretation of data.

Ethnic mobility and contextual effects are quite different components of category jumping, which result in people changing their ethnicity responses. Ethnic mobility refers to people changing how they identify their ethnicity over time. For example, the social environment of people may change in ways that lead them to identify themselves with additional or different ethnicities. The time frame varies depending on the underlying drivers and may reflect long-term processes resulting from changes in social environment or living arrangements, Partnership formation, change of job and moving to a different area are all examples of such changes. Conversely, changes may also occur over a shorter term. A typical example is where people might identify themselves differently at work and at home. Generally, people studying ethnic mobility focus on longer-term processes as these processes affect time series data. At an individual level, these changes generally relate to personal social changes but are not independent of the wider social context. Hence, underlying real-world societal changes cause wide-scale ethnic mobility, often over a relatively short period of time.

People may provide different responses depending upon the context or circumstances in which they are asked about their ethnicity identification eg, completing a self-administered form as opposed to answering an interviewer’s question. The context effect is quite distinct from ethnic mobility because it reflects the way people respond to how (the mode) and where/why (the circumstances) the information is collected. A common cause of change relates to the perceived purpose of the data , eg a person’s responses may differ where they understand the data in one collection relates to familial information and in another collection to social environments. This does not necessarily mean people are completing a form irresponsibly; rather, they may be providing ethnicity responses that best reflect how they identify themselves relative to what they understand to be the purpose of the information. Since many collections are not clearly statistical in function, perceptions of the purpose of ethnicity questions may include the indication of cultural needs, identifying cultural resources, and other positive, or indeed negative, discriminatory purposes, eg people may identify themselves differently when completing educational enrolments, benefit applications and census forms.

Collection of ethnicity information is by self-identification. However, self-identification is not always possible for a variety of reasons, eg in the collection of birth and death information and in certain cases of sickness. In such cases, proxy responses are collected. The effect of proxy responses is an issue, because in this case the individual has had their ethnicity identified by a third party on their behalf based on the third party’s perception of their ethnic identity rather than their own. This situation is most transparent in the collection of information on births and deaths. How ethnicity is recorded for births differs in a number of interesting ways from how it is recorded for deaths. At birth, ethnicity is supplied by parents. In making their decision, parents may consider factors such as ancestry, how the child will be brought up, and what degree of cultural competency the child will attain. In making this assessment, standards are often set relative to the parents’ or grandparents’ strength of ethnic affiliation. At death, ethnicity is supplied by next-of-kin. This is commonly a child, grandchild, spouse, sibling, parent or grandparent. While the same criteria may be applied for assessing ethnicity, a grandchild's judgement on the strength of ethnic affiliation, for example, is unlikely to match the criteria a grandparent would use.

One aspect of ethnic mobility which causes some concern is a perception that ethnic mobility involves losses to a group. However, this is not necessarily the case because ethnic mobility often involves not the loss of a particular ethnicity, but instead the acquisition of additional ethnicities. The result is that people who may previously have identified themselves as, for example, being of Tongan ethnicity begin to identify themselves as Tongan and Māori. In some cases there are losses to one or more groups, though the trend at present is towards multiple responses, so that gains tend to outnumber losses. However, generally, only net results of ethnic mobility and contextual effects can be easily derived. It is generally not possible to identify the gross flows between categories, limiting the ability to analyse in detail the magnitude of change.

Sources of data

Ethnicity is collected in a wide range of situations. Information on the ethnicities of people enumerated by a census, survey or administrative data source should be collected at the same time as other information collected and should, wherever possible, be supplied by the respondents themselves.

Survey data is collected by a number of means, such as face-to-face or telephone interviewing, electronic capture or postal questionnaires, and generally involves a carefully designed and selected subset of a subject population. The data collected is influenced by the question used, the mode of collection and the environment in which the data is requested. When designing surveys, these aspects need to be considered. In general, self-identification of ethnicity is required but in practice proxy responses may be the only feasible option. In the case of longitudinal surveys, ethnicity should ideally be collected at every phase, rather than the common practice of asking for ethnicity at first contact and carrying forward the original responses.

Administrative collections use a wide variety of forms and means of collection. Such collections are generally understood by respondents to have a specific purpose, a factor which influences how people respond to an ethnicity question. Questionnaire design in this case has a direct impact on the quality of the data and data consistency with other collections.

The objective of census data is to survey an entire population. That said, it is essentially similar to survey collections in the manner and type of information collected except that self-identification of ethnicity is specifically sought. The Census of Population and Dwellings is an important source of ethnicity data for small areas and small ethnic groups in New Zealand. In 2001, ethnicity was one of four core census variables (the other three being age, sex and location) identified as the foremost census variables. Resources were applied to these foremost variables to ensure that outputs were of the highest possible quality.

Collection issues: Data collection consistency

The Review of the Measurement of Ethnicity recommended that data resulting from collections should wherever possible be consistent across time and between collections. Therefore, the information collection process should endeavour to use questions and modes of enquiry that achieve such consistency. In some cases, altering the question may result in more consistent information than using the same question. For example, following the development of the 1996 Census questionnaire, the death registration forms adopted the new question. This change resulted in ethnicity data for deaths that was relatively consistent with the 1996 Census data and highly consistent with the 2001 Census data, correcting what had previously been a major problem in the reporting of mortality rates by ethnicity.

Collection issues: Reducing the number of multiple responses

It may not always be possible to retain all responses given, because many older data storage systems limit the number of responses that can be retained in a database. Nevertheless, keeping all responses should always be regarded as the ideal solution. This issue is more commonly faced in the collection of data, although it is also occasionally an issue in data output when the number of responses needs to be reduced. Where this is the case, reduction should only be done after careful consideration of the implications for the analysis of the data. One scenario where this may occur is when comparison is made between a collection that collected up to three ethnicities per person and another collection that collected a larger number of ethnicities. In such a case, it may be necessary to reduce the second data set to three responses for valid comparison. However, before doing so, it should be established that the methods used to collect the data were sufficiently similar to warrant taking such a step.

Where it is deemed essential to reduce the number of given responses to a lower maximum number of responses, a randomised selection should be made that is compatible with the methodology used for input as described in The Statistical Standard for Ethnicity, 2005. For detailed information about this methodology and its application, see *When Individual Responses Exceed Input Storage - A Procedure For Unbiased Reduction* available at www.stats.govt.nz.

Only records with more than the maximum number of responses will be affected. The primary principle behind this approach is conservation of information at Level 1 of The Standard Classification for Ethnicity, 2005. The Level 1 categories are European, Māori, Pacific Peoples, Asian, Middle Eastern/Latin American/African (MELAA) and Miscellaneous.

The steps involved are:

- remove any residual responses (first ‘Not Stated’, then the other residual responses such as ‘Don't Know’)
- remove less detailed or broad responses where there is a more specific one in the same Level 1 category (eg, remove Indian nfd if Gujarati is present)
- randomly remove specific responses within a Level 1 category until the required number is reached or there is only one response left in that category eg, at least one of Gujarati, Tamil, and Bengali should be kept because all three responses are classified within the Asian Level 1 category
- if reducing responses to six per person then all Level 1 information should be retained
- if reducing responses to three per person then Level 1 information may not be retained for all categories

It is very important to note that when (and only when) there is only one response remaining in any Level 1 category, should any Level 1 category be removed from the data, eg removing any response from people who have identified themselves as being of Scottish, Māori, Tongan, Thai, Somali, and New Zealander ethnicities will remove the Level 1 category that the ethnicity belongs to

Each step should only be carried out on records that have more than the required number of responses remaining after the application of the previous step.

Users of ethnicity data need to be aware of the implications of randomised reduction of the number of ethnicity responses, both where it has been used for data input as well as where it has been applied to output. Information is conserved so that all Level 1 categories are kept wherever possible. This is important because most analysis of ethnicity and analysis of related-policy is couched in terms of ethnicity at Level 1. Similarly, in cases where it is not possible to keep all Level 1 groups, random reduction of responses minimises the effect on the relativity of groups to each other and fairly represents the diversity and structure of the population. This contrasts with the previous prioritisation system where relativity between groups was frequently distorted. Special care is needed when comparing prioritised data with any other ethnicity data (including other prioritised data) for this reason.

Collection issues: Dealing with non-responses and residual codes

In many collections, responses are missing from some records and some responses given are coded to residual codes (eg 'Out of Scope'). Wherever it is necessary to derive information, such as the percentage of people who are of a particular ethnicity, the percentage should be calculated from the number of people with at least one specified response. At no time should totals which include non-responses be used to derive rates because doing so will under-report the derived rates (except in the extremely unlikely event that every non-respondent was not of the ethnicity being examined). Because non-response is known to occur at different frequencies for people of various ages, locations and ethnicities, the effect on the data differs between ethnic groups.

Data users are cautioned that some older published information may be based on data categorised as 'Non-Māori' where what is referred to as 'Non-Māori' was simply the count of people who did not state that they were of Māori ethnicity and frequently included all people who did not report an ethnicity. The problem with data of this type is that it assumes that all non-respondents were not of Māori ethnicity, which is clearly illogical. Data of this type is not comparable with current data based on specified totals. This data based on specified totals in turn assumes that the distribution of ethnicity among non-respondents is the same as for the specified population. Such an assumption is likewise invalid, but it is more fairly representative of the population than the previous approach.

Comparing collections

When data is compared across two or more different collections, consideration needs to be given to how and when the data was collected. In some cases, the available data may not be directly comparable, while in others, the most appropriate subject population may not be available. In this situation, assumptions must be made (and identified) about uncertainties in the resulting data. For example, prior to 1996, death registrations recorded only whether the deceased were of Māori or Pacific ethnicity. There was no measure of deaths by ethnicity for people of other ethnicities because this information was not collected. It also reported only partial ethnicity data and did so in a way that was incompatible with other data sources. Of particular relevance, because of the prominence given at the time to differences between Māori and non-Māori mortality, is that the data collection method on death registration forms provided no valid means to measure non-Māori mortality. The actual level of non-responses in this case is unknown. It is similarly unsafe to assume that all Māori or Pacific deaths were recorded correctly. This becomes especially critical when measures (such as mortality rates) are derived from a base population measured in a different way. In this particular case, the subject population was taken from the population census or estimates derived from the census which collected and output ethnicity in a different way.

As a brief checklist for analysing differences between collections, the following are representative of the issues to be considered:

- collection method
- question changes
- number of responses collected per respondent
- how the responses have been processed
- input issues
- output issues.

One of the major issues addressed in the Review of the Measurement of Ethnicity was the incompatibility of data from different sources. Many historical data sources recorded ethnicity in a variety of ways and different sources often used different questions or processing methods. Consequently, it is not always easy to work out valid methods for comparing sources. Because so much policy analysis requires trend analysis, a good understanding of changes in collections over time is essential.

When data is drawn from more than one source to be integrated or compared, it is important to consider the effect of the different ways in which the data has been collected and how it has been output. This is particularly important in cases where the sources collected different numbers of responses or where the data has been tabulated in different ways, especially if rates are being derived. For example, a collection that contains just one ethnicity response is likely to have used some method of prioritising data to remove excess responses. Such methods have a biasing effect on the data with the consequence that conclusions based on the data may be highly misleading. In the case of people of Māori and Pacific ethnicities, 23 percent of people of Pacific ethnicity under the age of 15 years also identified as Māori in the 2001 Census. The systematic prioritisation of the data, used in the late 1980s and early 1990s, which gave highest priority to Māori, for example, excluded people from the Pacific count simply because they happened to also identify themselves as Māori. This produced misleading results that under-represented people of Pacific ethnicities.

A particular problem arises where administrative data (eg police statistics) is used as a numerator and survey data (eg the census) is used as the denominator to derive rates. In such cases, not only is the data collected in different ways, but the underlying concepts may also differ. The ethnicity recorded for offenders may represent physiological appearance as perceived by an arresting officer whereas ethnicity in the census is recorded according to self-identified socio-cultural affiliation. The police record is also less likely to include multiple responses than the census data, thereby providing an example where the purposes of two collections differ.

Output choices

There are currently two principal recommended approaches to ethnicity data output and analysis: total responses and single/combination outputs. In this section we describe both approaches and discuss some aspects of their use and limitations.

Total response output

Total response ethnicity data should be used wherever possible. This output method counts every ethnic group that a person identifies with. This is consistent with how people report their identity in so far as people of two or more different ethnicities consider themselves to be members of each of the identified ethnicities. People with responses which fall into more than one group are counted once in each group with which they identified themselves. For example, people of Samoan, Tongan and German ethnicities would be counted (when outputting at the highest level of the classification) once in the Pacific category and once in the European one. This means that the sum of the ethnic groups will be greater than the number of people. This is similar to many other commonly used variables, such as income sources or iwi affiliations, in which the sum of the categories is greater than the count of people.

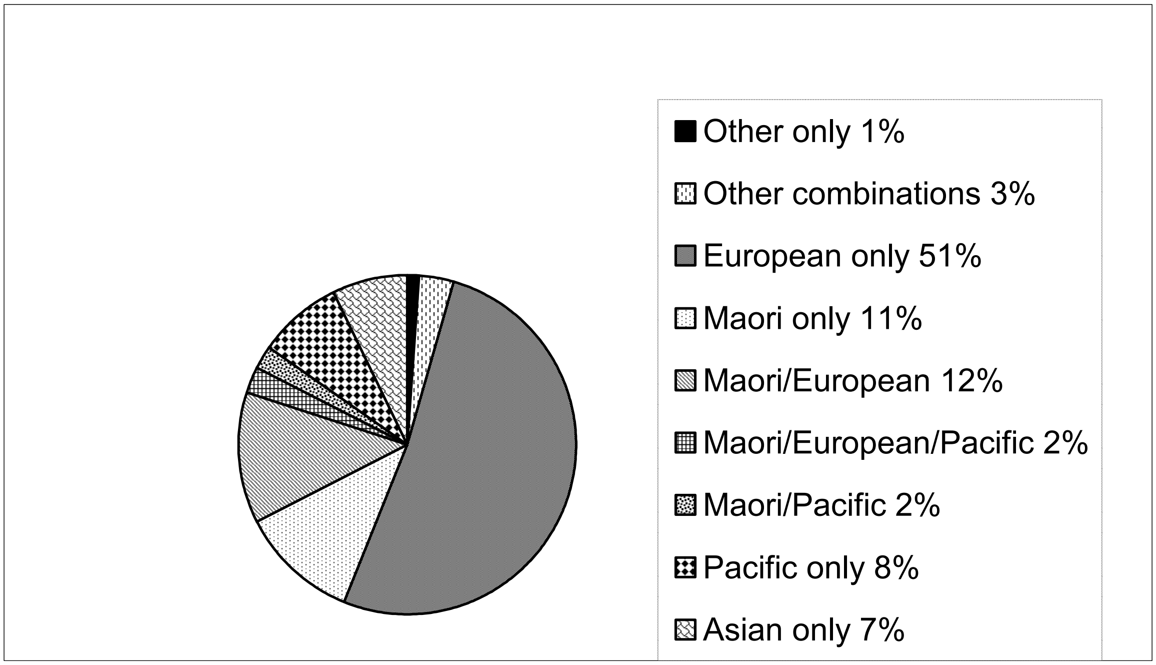
The advantage of total responses is that the relative size of the groups within the population is fairly represented (using as the denominator only the count of people for whom ethnicity is available). The proportion of a group which overlaps with other groups may be large and reflects social diversity. The overlap however is only observable at a high level and in order to gain an understanding of such overlapping, single/combination analysis may be required.

Single/combination output

Single/combination ethnicity data provides much more detailed information. It also has the advantage that the sum of the categories is the same as the count of people who specified ethnicity. The overlaps between ethnic groups is often significant. Any analysis of a population should include an understanding of the relationship between ethnic groups and such overlaps. For example, people who identify themselves with two groups may possess different characteristics from people who belong to one of those groups but not the other. Therefore, when used in conjunction with total response data, single/combination data provides useful information on the components of ethnic groups. This is a very powerful method for identifying diversity and dynamics within and between groups, and it has the potential to assist in explaining trends. See Figure 1 for an example of single/combination output.

The disadvantage with single/combination data lies in the large number of potential categories. Even at Level 1 where there are only six ethnicity groupings (European, Māori, Asian, Pacific, Middle Eastern/Latin American/African, Miscellaneous), there are 61 possible categories (ie each combination plus Not Stated), although some combinations would be very small or empty, and would therefore be collapsed. While not all combinations would be of interest, care is needed to ensure that groups chosen for analysis are consistent with the limitations imposed by the quality of the data. Analysis of ethnicity data (eg infant deaths versus births, comparing intercensal population change, births and deaths versus census) suggests that single/combination data is currently far less stable than total response data. Single/combination data appears to be very sensitive to ethnic mobility and contextual effects because these are frequently transitional processes which occur across boundaries between combinations of ethnicities within groups.

Figure 1: Distribution of live births by ethnicity of child in single/combination groups, year ended 31 December 2003



New Zealander Output

'New Zealander', 'Kiwi' and other similar responses were formerly categorised in the Level 1 European group under New Zealand European responses. They are now categorised in the Level 1 Miscellaneous group under New Zealander to reflect their growth and the lack of evidence to support the previous approach. This means that the count of the New Zealand European category and the European group for the 2006 Census of Population and Dwellings will not be directly comparable with previous censuses. A similar issue applies to all other collections.

If time series data is required, the comparability of the European ethnicity grouping with previous periods requires special treatment. Collapsing the Level 1 Miscellaneous group into the Level 1 European group will provide a reasonable approximation of the former Level 1 European category. ‘New Zealander’ responses will account for the vast majority of people in the Miscellaneous group. In the past very few people have given ‘New Zealander’ as a response at the same time as any ethnicities within the European ethnicity grouping. While a few people may provide responses in both the European and the Miscellaneous groups, and strictly should be counted only once in the combined group, any error arising from simple addition of the ‘New Zealander’ count to the European group will be much smaller than other errors. However, in any collection where the other ethnicities in the Miscellaneous group are significant it would be preferable to add the ‘New Zealander’ responses only to the European grouping and retain the other ethnicities under Miscellaneous.

However, when representing the new Miscellaneous group in a time series, it may be preferable to simply introduce the group to the analysis from the point in time when it is output, rather than attempting to 'turn back the clock'. This approach would produce a clear break in any time series and would require noting at the time.

Both approaches, reconstituting groups and showing a break in a series, have strengths and weaknesses and are something that users need to consider on a case by case basis. Statistics New Zealand provides concordances which match old ethnicity codes to new ethnicity codes for use when converting data collected under the old classification to provide the best fit with data collected under the new classification.

Managing Data

Managing data refers here to the process of getting data into a form which enables valid analyses to be made. This section should be read in conjunction with the section on collection issues with ethnicity data and handling incompatibilities between collections.

Perhaps the most important issue is whether using ethnicity as a frame of reference is appropriate in the particular situation under consideration. The environment which the data purports to reflect should be what we are trying to determine. Since much policy is differentiated by ethnicity, analysis by ethnicity is a common starting point. It is also common to assume that this is a primary causative parameter. There is a risk that a cause or trend ascribed to ethnic diversity may be driven by another primary factor such as age.

The main focus of interest in the analysis of ethnicity is often the differing migration and socio-economic histories of groups relative to other groups, although the demographic features being analysed are often quite age-sex specific. The age profiles of ethnic groups may differ significantly (especially for people of multiple ethnicities). Great care is needed when a characteristic of one group as a whole is compared with another group as a whole. Standardisation for age or some other key variable is an important method of adjusting in order to take this into account. For example, while individual age groups may be compared directly, the different age and sex structures of ethnic groups mean that, before populations as a whole can be validly compared, the relevant groups should be age-standardised. For further information on standardisation see http://www.population.govt.nz/glossary/p-z.htm#standard.

When analysing ethnicity data, it is preferable to compare groups to the total number of people with at least one valid response. Not-specified cases are by no means insignificant in some collections, and it is unwise to assume that in every case the characteristics of this group are similar to the people who did specify their ethnicity.

Use of indicators

One approach is to include in data sets separate derived variables which indicate whether or not a person identifies with target ethnicities (eg a Pacific indicator would identify whether or not a person had specified that they were of one or more Pacific ethnicities). These ethnicity indicators should be included in (or added to) data sets where possible because they provide a useful tool for data analysts. People can easily filter data to obtain information on particular Level 1 populations (eg all people of Asian ethnicities). In addition, when used in combination, indicators enable rapid and consistent extraction of information on particular overlapping groups (eg people of both European and Pacific ethnicities). Indicators should, of course, supplement rather than replace individual ethnicity variables.

Building indicators should be appropriate to the analytical context of the data. There are a number of options, but it is helpful if the categories can be aggregated easily and that categories be mutually exclusive. Generally, indicators would be included for Level 1 ethnic groups only, though indicators for other levels and individual ethnicities may be needed in particular situations.

The minimum number of categories for an ethnicity indicator is three (people of ethnicities in X ethnic group, people with specified ethnicities but none in the X ethnic group, and people for whom no ethnicity has been specified). It is incorrect to combine the last two categories. The categories in list form are:

- X ethnic group
- not in X ethnic group, but at least one valid ethnicity response
- ethnicity not available.

A more versatile indicator might include four categories. In this case the first category (people of ethnicities in X ethnic group) might be split into two categories: (1) people of ethnicities only in X ethnic group, and (2) people of ethnicities in X ethnic group as well as of ethnicities in some other ethnic group. The other two categories (people with specified ethnicities but none in the X ethnic group, and people for whom no ethnicity has been specified) would remain the same. The categories in list form are:

- X ethnic group only
- X ethnic group and in other ethnic group(s)
- not in X ethnic group, but at least one valid ethnicity response
- ethnicity not available.

Other indicators may be needed for special purposes, but generally combinations of indicators can be used together to get sets of combinations as required. For example, using Pacific and Māori indicators together can readily provide information on people of both Māori and Pacific ethnicities or people of Pacific but not Māori ethnicity.

Using indicators in conjunction with total response data is particularly powerful. This enables analysis of the composition and internal diversity of major groups, eg an analysis of ethnic diversity among people of selected ethnicities of interest. This can be readily done, without the need for complex recoding, by cross-tabulating the ethnicity indicators with the total response data.

The use of indicators to define a non-X ethnic group is strongly discouraged. The principal reason for this is that people are generally asked which ethnicities they positively identify themselves with. It is commonly assumed, wrongly, that failure to tick the 'Māori' box is deliberate - as though people had actually been asked two questions: Are you of Māori ethnicity? and Are you non-Māori? People are rarely asked which ethnicities they do not identify themselves with. Therefore, any analysis of people who have not identified themselves with a specific group cannot be assumed to represent people who do not belong to that group. Absence of a particular response does not necessarily equate to absence of that characteristic; it may simply be the way that person answered in the particular context. This can be especially problematic given the high level of interest of people in such dichotomous analyses, eg in disparity analysis, and as such, it can be very easy to arrive at misleading conclusions.

In particular, analysis should not treat the non-Māori group as though it were a valid ethnic group. The demographic processes which shape the population defined as "non-Māori" are not the same as those which shape a true ethnic group. Moreover, more than half of all people of Māori ethnicity also identify themselves with other ethnicities and this is likely to continue to increase due to miscegenation. Half of all Māori babies are born to at least one non-Māori parent (similarly 40 percent of Pacific babies now have a non-Pacific parent). It is recommended that measures of demographic or socio-economic performance of a group be made against the whole population with specified ethnicity rather than with the fictitious non-group.

Time series

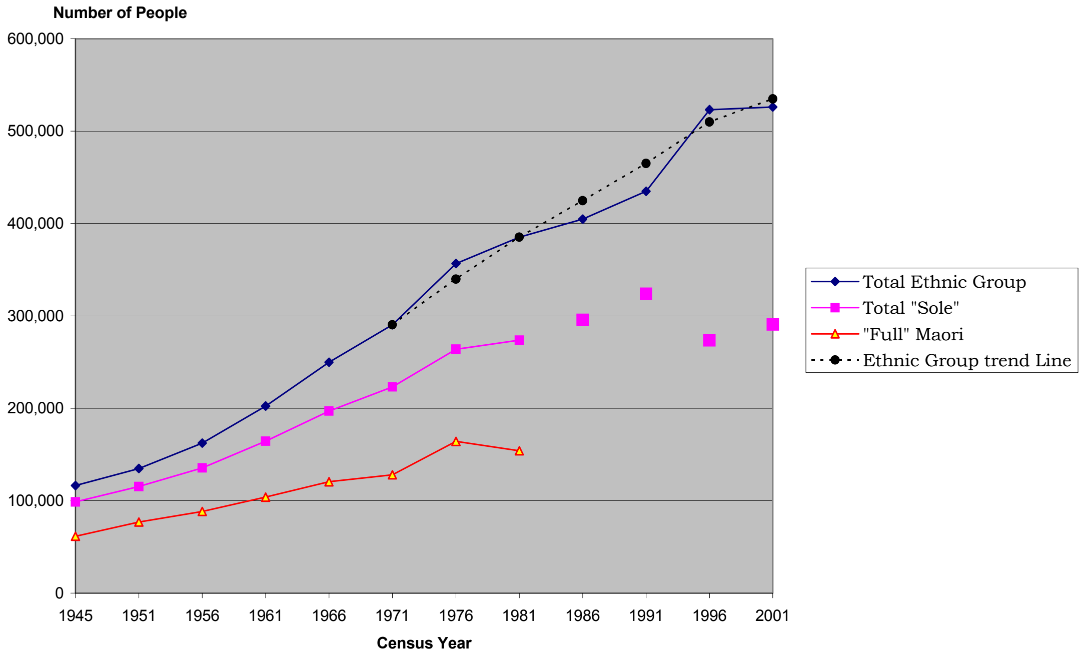
Policy analysts and social scientists frequently need to understand how population characteristics have changed or will change over time, in order to contextualise the current state of the population. The principal tool for doing this is a time series. This section discusses issues related to time series data derived from the same source, eg the census or regular surveys. Although longitudinal surveys also present difficulties where ethnicities are not collected at each point in time, these are not specifically included in the discussion here due to there being a simple mechanism for understanding the changes: these surveys usually contain essentially the same people in their sample throughout the life of the survey and it is simply a matter of asking for the ethnicities of respondents each time.

Time series analysis is a central component of policy planning, yet the collection of ethnicity changes over time for every data source and both the level of coverage and quality of data becomes an issue at each data point. It is important to consider the effects of these changes in the definition, collection and coding of the data. This is particularly important for Māori data, which has been collated in many different ways over time. Constructing a valid time series based on fluid concepts such as ethnicity requires careful consideration of how the concept has changed over time, how the collection and processing of data has affected the pattern and a range of data quality issues. Real world changes in patterns of identity do occur and these should be considered both in the construction of time series and in the forecasting or backcasting of trends.

Looking at Māori ethnic group counts in the census offers a practical illustration of some of the issues involved in deriving time series for ethnicity. Māori ethnic group counts serve as a good example because of the importance of this group in policy analyses. It should be noted, however, that this exercise is purely illustrative of the issues involved and similar issues are associated with other data sources and people of other ethnicities.

Partly because of the way ethnicity information has been collected in different censuses, with different questions being used and different concepts being applied to the data, significant differences have appeared in the data in recent years. Figure 2, below, which plots changes in the Māori population from the 1945 Census to the 2001 Census, illustrates this point. It is immediately apparent that there is a reasonable fit between the data collected since 1996 and the data collected prior to 1981, while care needs to be taken when comparing recent data with data from either the 1986 or 1991 Census. The term "sole" is used here in the obsolete sense found extensively in the historical data sources, and refers to people who identified themselves as ‘half or more Māori’ up to 1981 or ticked only the ‘Māori’ tick box on the census forms from 1986 onwards.

Figure 2: Changes in Māori census populations 1945-2001



The level of compatibility between 1996 and 2001 can be seen when changes in age structure are considered within the proportion of the Māori population with one ethnicity response (previously called 'Sole-Māori' population). To compare the age structures, Figure 3 plots the census populations from 1981 to 2001 by the age people would have been in 2001, effectively comparing birth cohorts. Although there are structural differences associated with migration, ethnic mobility, under-enumeration and mortality, it is clear that a radical shift took place between 1991 and 1996 in the way people identified themselves. This reflects a significant real-world change among people of Māori ethnicity towards multiple responses between 1991 and 1996.

Figure 3: Single Māori response 1981-2001 Censuses, age adjusted to 2001 notional age

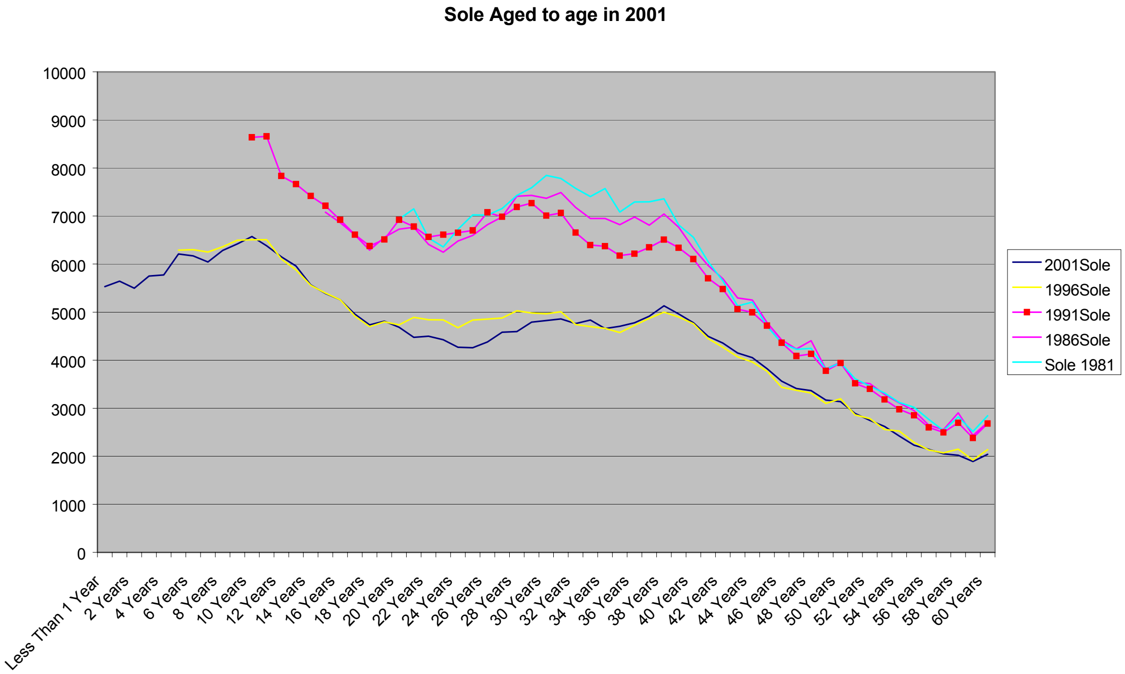
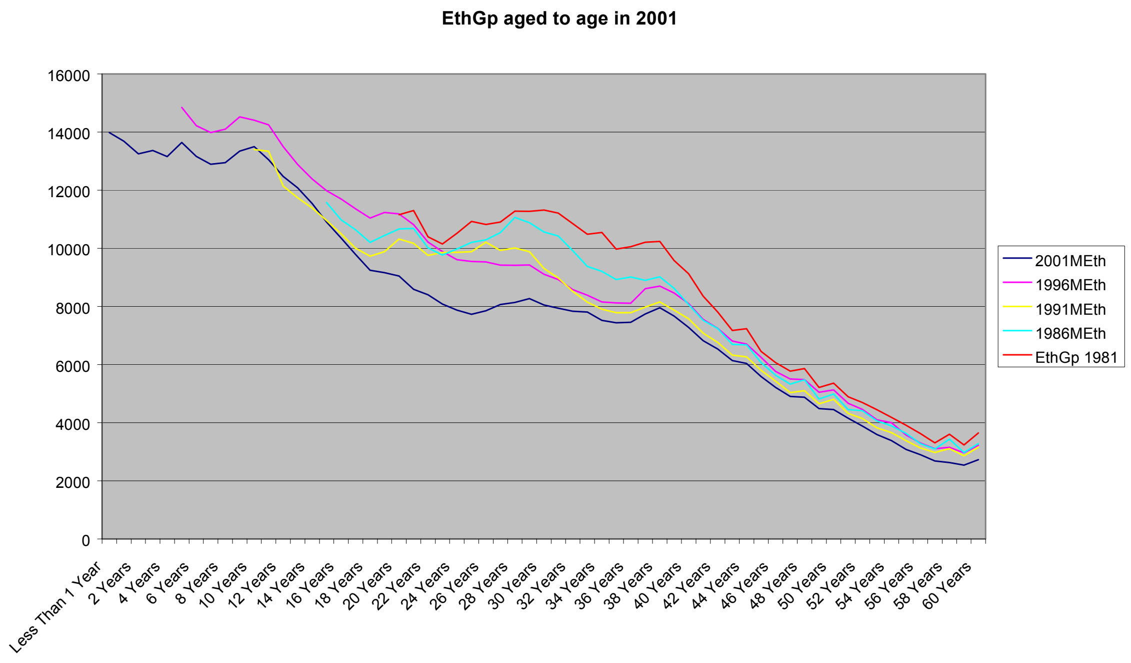


Figure 4 shows that this is a feature of the internal structure of the Māori ethnic group rather than the group as a whole. However, the changes in the ethnic group over time illustrate a number of aspects which make the creation of a sound time series problematic. The net gains and losses to the group due to the combined effect of emigration, return migration, mortality, fertility changes, underenumeration, and ethnic mobility can be seen among particular age groups.

Figure 4: Māori ethnic group 1981-2001 Censuses, age adjusted to 2001 notional age



These issues also affect people of other ethnicities. For smaller ethnic groups and ethnic groups undergoing rapid change, the magnitude of the problem is larger. Therefore it is rarely possible to derive a robust time series for any other group, except for the most recent period or into the near future. In censuses and surveys, there are large differences between levels of undercoverage and non-responses across different ethnic groups which understate some ethnic groups significantly. These differences vary over time in major ways. For example, as Table 1 illustrates, the population estimates indicate that, while New Zealand’s resident population was 4 percent larger than that enumerated by the 2001 Census, the Māori ethnic group population resident in New Zealand was 11 percent larger; the Pacific population was 13 percent larger and the Asian population 14 percent larger. The larger differences seen among people of Asian ethnicities reflect, in part, higher levels of net migration gain between the census date in March 2001 and the June 2001 estimates, while the differences for people of Māori and Pacific ethnicities are largely due to census undercount. In each case there were wide variations across age groups, with some ethnicities understated by 20 percent or more at some ages. The estimates, which are adjusted for these variations, provide the correct denominators against which to interpret contemporary demographic events because the subject population includes those people not included in census counts. Wherever possible, official population estimates should be used as the base population for current and past populations, while official population projections should be used for future change.

Table 1: Comparison of census and estimated populations

Age group	New Zealand Total		Ethnic Group							
			European		Maori		Asian		Pacific	
	Census	Estimated	Census	Estimated	Census	Estimated	Census	Estimated	Census	Estimated
0-4 years	270,801	281,000	195,180	207,800	67,560	74,600	18,375	20,400	32,775	36,600
5-9 years	286,200	295,500	207,687	219,600	66,114	72,700	18,381	20,000	30,483	33,600
10-14 years	290,739	300,700	213,057	225,100	62,808	68,900	19,521	21,100	26,889	29,800
15-19 years	265,281	277,600	187,872	201,600	49,530	55,200	27,102	30,200	21,486	24,100
20-24 years	239,784	255,600	167,382	183,100	42,093	48,300	24,018	28,900	19,779	22,800
25-29 years	246,900	257,200	180,117	192,800	40,164	45,400	17,979	21,700	17,976	20,800
30-34 years	279,276	291,600	206,802	223,200	39,252	44,300	19,947	23,100	17,778	20,100
35-39 years	297,462	306,700	222,828	236,800	38,322	42,400	22,722	25,700	16,011	18,200
40-44 years	285,618	296,900	220,077	236,400	32,859	36,500	19,806	22,400	12,753	14,500
45-49 years	251,784	261,700	198,456	212,700	25,095	28,200	15,690	18,200	10,134	11,600
50-54 years	236,169	245,100	193,026	207,100	19,473	21,800	11,520	13,600	7,977	9,100
55-59 years	182,262	189,300	151,572	162,900	13,827	15,300	7,161	8,500	5,667	6,600
60-64 years	154,569	161,000	128,352	139,100	11,550	12,800	6,192	7,200	4,461	5,100
65-69 years	127,914	130,900	108,705	115,300	7,938	8,800	4,431	5,200	3,147	3,700
70-74 years	118,257	120,800	104,412	111,100	5,070	5,600	2,592	3,000	2,244	2,500
75-79 years	94,506	96,600	85,434	91,200	2,688	3,100	1,485	1,700	1,269	1,500
80-84 years	61,110	62,500	55,728	59,900	1,215	1,400	753	900	636	800
85 and over	48,639	49,800	44,748	48,300	726	800	501	600	339	400
Total	3,737,277	3,880,500	2,871,432	3,074,000	526,281	586,000	238,176	272,400	231,798	261,800
Percent difference		3.8		7.1		11.3		14.4		12.9

Source; usually resident population, 2001 Census, estimated resident population at 30 June 2001

For historical populations and for enquiries requiring extensive socio-economic analysis, the Census of Population and Dwellings is the most comprehensive source of information. The census forms the basis of both population estimates and population projections, and remains the primary source of social and economic information on population. For ethnicity, the census is the only comprehensive source of detailed information on ethnicity because population estimates and projections of ethnic groups are only available for some Level 1 categories. For analysis of ethnicities other than at Level 1 and for smaller Level 1 groups, the census should be used. Where the current population at a more-detailed level is critical for a particular purpose, eg for observing a targeted short-term health monitoring programme, ad hoc estimates may be derived from the census.

When output becomes available from the 2006 Census of Population and Dwellings, the current intention is for it to follow the format shown in Table 2 below. This reflects the changes in the methods used to collect data. Because the questions used in 2001 and 2006 will be the same, and the data collected in 2001 is closely comparable with 1996, and to a lesser extent with 1991, this will provide the best available basis for the analysis of ethnicity over time.

Table 2: Changes in methods used to collect data

1991	**1996**	**2001**	**2006**
up to 3 responses (prioritised at input)	up to 3 responses (prioritised at input)	up to 6 responses (prioritised at input)	up to 6 responses (randomised after input)

Data integration: Issues

Analysis of ethnicity data often involves comparing one population or data source with another, or interpreting change over time. A particular problem is defining which is the appropriate subject population to use as the basis for the analysis. For example, the decision on data source devolves into whether the focus of interest relates to past, present or future populations or on change over time, and what the scope of the enquiry is.

For current populations, the official population estimates provide the most up-to-date and complete source of information. These estimates are based on census populations but are adjusted for net undercount, people temporarily overseas at the time of the census, and non-responses to ethnicity together with births, deaths and permanent/long-term migration between the census and the base date for the estimates (30 June of any year).

Any analysis of a population should include an understanding of people of more than one ethnicity who therefore belong in more than one ethnic group. Comparison of ethnicity combinations can be made to the census when considering the representativeness of sub-populations in sample surveys. This will provide some information on the comparability of the data sources. Combinations of ethnicities demonstrate that ethnic affiliation is not a singular experience and they underline the complexity of this type of information.

Particular care is needed with derived measures which involve using single/combination ethnicity data from different collections. Different collections reflect multiple ethnicities in different ways. For example, mortality measures involve data from death registrations, which collect ethnicity by proxy, and data from the population at risk, which is based on population estimates derived from the census where ethnicity is generally self-identified. Among the consequences, infant mortality for the single-response Māori population is currently overstated by 20 percent whereas this is not the case for the Māori ethnic group as a whole. This example demonstrates why the use of concepts like ‘Sole-Māori’ are not recommended except in the context of the internal structure of an ethnic group as a whole within a single collection. Applying total ethnic group data responses provides a more robust measure across multiple collections.

Data integration: Procedure

What should be done when two data sets contain different information depends on the particular situation. It is not necessarily valid to simply take the most recent response. No general rule can be given on how to handle each case, except that it is frequently better to take the response associated with the denominator in any rate calculations. This can best be explained by way of some examples.

All rates are ultimately calculated by dividing the number of cases (the numerator) by the number of people at risk of whatever is being discussed (the denominator).

$$\text{Rate} = \frac{\text{Numerator}}{\text{Denominator}}$$

Example using birth statistics

The derived rate of births per woman is calculated from the number of births divided by the number of women of childbearing age. This requires knowing how many births occur in the period concerned and also how many women could have given birth (referred to as the population at risk). Even for the total population this is problematic, because the estimated population changes throughout the period and there are a range of uncertainties which need to be accounted for. For ethnic groups, this becomes more difficult when dealing with pre-1996 births, because the way ethnicity was collected on birth registrations differed from how it was collected in the census. This is important because the census data formed the key information source for estimating the population at risk. It should be noted, however, that in this particular instance the data from 1996 to 2004 is closely comparable.

Example using crime statistics

In the case of crime statistics, the police may base their identification of people entirely on how they interpret the appearance of an individual, eg Asian, or Māori, or Pacific/Māori. However, people identified in this way may have identified themselves entirely differently in the census. Therefore, when the police counts are divided by the population at risk, the resulting rates may not be as reliable as expected, especially when dealing with small numbers of incidents. To demonstrate the magnitude of this issue, assume a situation where 20 males aged 20-24 who have been arrested for a particular group of offences are recorded as Māori offenders. A further 15 are recorded as Pacific offenders, and 10 are recorded as Asian/European ones. If the male population aged 20-24 living in the relevant areas numbered 15,000 (of whom 10,000 were of European ethnicity, 2,000 Asian, 2,000 Pacific and 2,000 Māori, always remembering that people can belong to more than one ethnic group), the following results will emerge: among people of European ethnicities the offending rate would appear to be 1 in 1,000; for people of Asian ethnicities it would be 5 per 1,000; for people of Pacific ethnicities, 7.5 per 1,000; and for people of Māori ethnicity, 10 per 1,000. However, five people among the Pacific offenders may have identified themselves as Māori/Pacific in the census, while 10 of the Māori may have identified themselves as Māori/European. Were the arrest record accurately to reflect the self-identified ethnicities, the rates would become: European, 2 per 1,000; Asian, 5 per 1,000; Pacific, 7.5 per 1,000; and Māori, 12.5 per 1,000. Variations of this magnitude have immediate consequences for any conclusions derived from the data.

Key messages

Wherever possible, analysis should be based on total responses for each group since this reflects both the number of people in the group and the relative proportion of the population that the group in question represents. The total of all groups will generally exceed the number of people in the population because people may identify themselves with more than one ethnicity. Total response counts should be output routinely.

People increasingly identify with several ethnicities. Where detailed information on the composition of the group and the inter-relationship between groups is required, analysis may include the use of data on combinations of ethnicities. Single and combination data should be made available wherever possible, taking account of the size of the populations in the sample and other relevant issues such as confidentiality constraints. The associated total response counts should always be output with the single and combination data.

As with any other information on the characteristics of people, data on ethnicity is subject to a range of factors which reflect collection and collation procedures and which have an impact on the information produced. Ethnicity is not fixed and it may change over time at both the individual and/or societal levels. Such changes are part of the nature of ethnicity.

References

Statistics New Zealand (2004). *Report of the Review of the Measurement of Ethnicity*, Statistics New Zealand, Wellington.

Statistics New Zealand (2005). *The Standard Classification for Ethnicity 2005*, Statistics New Zealand, Wellington.

Statistics New Zealand (2005). *The Statistical Standard for Ethnicity 2005*, Statistics New Zealand, Wellington.

Statistics New Zealand (2005). *When Individual Responses Exceed Input Storage - A Procedure For Unbiased Reduction*, Statistics New Zealand, Wellington.