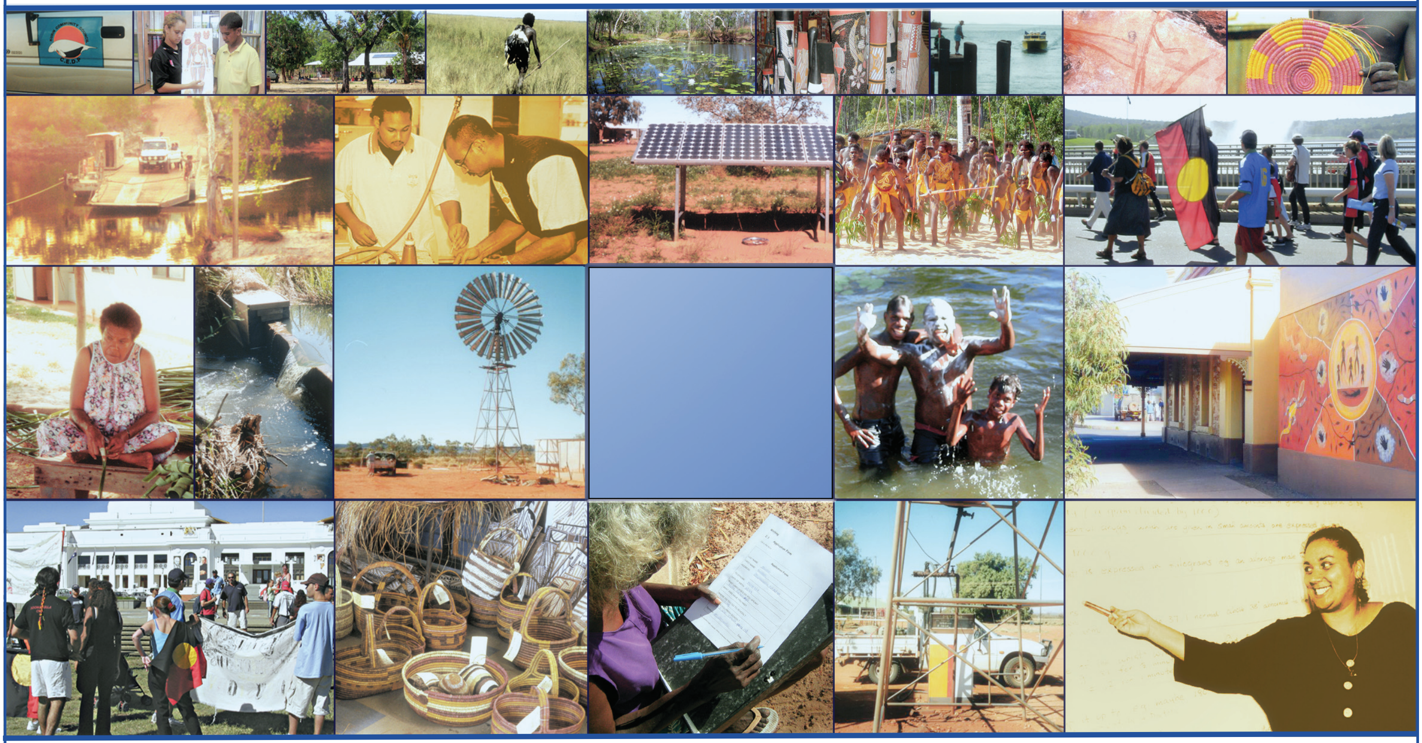
CENTRE FOR ABORIGINAL ECONOMIC
POLICY RESEARCH

THE AUSTRALIAN NATIONAL UNIVERSITY

Some Reflections on the Quality of Administrative Data for Indigenous Australians: The Importance of Knowing Something about the Unknown(s)

B.H. Hunter and A. Ayyar

CAEPR WORKING PAPER No. 51/2009



ANU COLLEGE OF ARTS & SOCIAL SCIENCES

SERIES NOTE

The Centre for Aboriginal Economic Policy Research (CAEPR) was established at The Australian National University (ANU) in April 1990. From 1990 to 2003 the Centre's main research partner was the Aboriginal and Torres Strait Islander Commission. Since 1 January 1999, CAEPR has operated as a University Centre and is currently funded from a variety of sources including the ANU, Australian Research Council, industry partners, and the Department of Families, Community Services and Indigenous Affairs.

CAEPR's principal objective is to undertake high-quality, independent research that will assist in furthering the social and economic development and empowerment of Aboriginal and Torres Strait Islander people throughout Australia. Its aim is to be a world-class centre undertaking research on Indigenous economic development issues that combines academic excellence with policy relevance and realism.

CAEPR is currently Australia's major dedicated research centre focusing on Indigenous economic and social policy from a national perspective. The Centre's publications, which include the CAEPR Working Paper series established in 1999, aim to report on Indigenous circumstance, inform public debate, examine government policy, and influence policy formulation.

Working Papers are often work-in-progress reports and are produced for rapid distribution to enable widespread discussion and comment. They are available in electronic format only for free download from CAEPR's website:

<www.anu.edu.au/caepr/>

Enquiries may be directed to:

The Centre for Aboriginal Economic Policy Research
Hanna Neumann Building #21
The Australian National University
Canberra ACT 0200

Telephone 02–6125 8211
Facsimile 02–6125 9730

As with all CAEPR publications, the views expressed in this Working Paper are those of the author(s) and do not reflect any official CAEPR position.

Professor Jon Altman
Director, CAEPR
College of Arts & Social Sciences
The Australian National University
March 2009

Cover page images courtesy of the Australian Institute of Aboriginal and Torres Strait Islander Studies and CAEPR staff members.

Some reflections on the quality of administrative data for Indigenous Australians: The importance of knowing something about the unknown(s)

B.H. Hunter and A. Ayyar

Boyd Hunter is a Senior Fellow and Aarthi Ayyar is a Research Assistant at the Centre for Aboriginal Economic Policy Research, College of Arts and Social Sciences, The Australian National University.

Centre for Aboriginal Economic Policy Research, Working Paper No. 51/2009
ISSN 1442-3871 ISBN 0 7315 4950 3
An electronic publication downloaded from <http://www.anu.edu.au/caepr/>.

ABSTRACT

The Repeat Offenders Database, which has been collated by the New South Wales Bureau of Crime Statistics and Research, offers a unique opportunity to analyse data quality issues for an important source of administrative data for Indigenous people. This paper provides several independent estimates of the population of Indigenous offenders by estimating the number of people with unknown Indigenous status who are likely to be identified as Indigenous in other circumstances. The main finding is that the Indigenous population of offenders are substantially undercounted in administrative data collections. The failure to account for this will understate the 'gap' between Indigenous and non-Indigenous outcomes.

Keywords: Administrative data, data quality, crime, ethnic mobility, Indigenous disadvantage, closing the gaps

ACKNOWLEDGEMENTS

This paper was written as part of an Australian Research Council Linkage grant (LP0776958) entitled 'An inter-disciplinary analysis of the dynamics of Aboriginal interactions with the criminal justice system' that involves the collaboration between the Australian National University and New South Wales Attorney General's Department. We are indebted to Jon Altman, Nicholas Biddle, Steve Moffatt, Tim Rowse, Kate Sullivan, and Don Weatherburn for comments on an earlier draft of this paper. Our thanks also go to Hilary Bek and John Hughes at CAEPR for editing and layout. Any errors are of course our own.

CONTENTS

TABLES AND FIGURES

INTRODUCTION

Why is the accuracy of administrative data for Indigenous populations an important question for researchers to address? If governments do not know exactly who is Indigenous then they cannot plan adequately to address Indigenous disadvantage. Consequently the major thrusts of Indigenous policy for 'Overcoming Indigenous Disadvantage' (OID) and 'Closing the Gaps' between Indigenous and non-Indigenous life expectancy are undermined (Steering Committee for the Review of Government Service Provision (SCRGSP 2003, 2005, 2007; Altman, Biddle & Hunter 2008).

OID: Overcoming Indigenous Disadvantage

ABS: Australian Bureau of Statistics

CGC: Commonwealth Grants Commission

ERP: Estimated Residential Population

Undercounts of Indigenous people are present in almost all data sets, even in the census, and hence it is difficult to know the exact number of people covered by a particular policy. The Australian Bureau of Statistics (ABS) uses sophisticated techniques to correct for this tendency towards under-enumeration in the Indigenous population in census data, but it is relatively rare to find similar corrections applied to data collected for administrative purposes by government agencies. In addition to undermining the credibility of the information provided from such sources, poor quality data have profound consequences for policy settings. For example, the Commonwealth Grants Commission's (CGC's) horizontal fiscal equalisation formula is partially based on the Indigenous Estimated Residential Population (ERP): hence data quality for Indigenous mortality and fertility will fundamentally affect the distribution of resources in Australia's federal system.

While data quality is a rather dry topic, it is clearly an important one. The objective of this paper is to present some analysis illustrating how one might go about critically examining administrative data on the Indigenous population. While many administrative data sets collect information on the Indigenous status of clients, often the quality of such data is questionable.

One of the largest categories of census response is the Indigenous status listed as 'unknown'. The fundamental premise of this paper is that it is important to understand this category lest the data quality be fundamentally compromised. If the unknowns are a substantial component of the population, then one cannot be certain one has correctly estimated the incidence of phenomena in the Indigenous and the residual Australian population. That is, if the unknowns are more like the non-Indigenous population than Indigenous Australians, then a comparison between the known Indigenous and non-Indigenous populations would overstate Indigenous disadvantage. Of course the obverse of this proposition is also true (i.e., if the unknowns are more like those who clearly indicate their indigeneity, then Indigenous disadvantage is likely to be understated).

It is obviously important to understand who identifies as an Indigenous person at a particular point in time, but it is arguably even more important to understand how someone's identity might change over time. This is because, as alluded to above, administrative data are increasingly used to ascertain whether Indigenous outcomes are improving according to various indictors (e.g. in OID reports). If data quality is unreliable and uncertain, then one should question the level of investment warranted to achieve some desired outcomes. Often there are only subtle changes in outcomes and indicators, which would probably not be significant if the true measure of the reliability of estimators were used. If real resources are diverted from the programs that worked because of spurious trends in outcomes, this will have set back the nominal goal of overcoming Indigenous disadvantage.

ROD: Repeat Offender Database

BOCSAR: Bureau of Crime Statistics and Research

This paper illustrates some of the issues involved in using administrative data from the Repeat Offender Database (ROD) provided by the New South Wales Bureau of Crime Statistics and Research (BOCSAR) (Snowball & Weatherburn 2006). This set has several features that make it suitable for a study of the quality of Indigenous data. First, it is collected over a period of time during which it is possible for Indigenous people to change their identification; and second, data on Indigenous status is collected from two independent sources (the Department of Corrective Services and the New South Wales Police) and this allows us to validate our estimates.

PES:
Post-Enumeration Survey

DSE:
Dual System Estimator

The standard approach adopted to estimate the number of people missing from any particular enumeration is to conduct a follow-up survey (Marks et al. 1974). Such a survey is undertaken after major censuses in most developed countries; in Australia it is known as the Post-Enumeration Survey (PES).[1] Another method for estimating the potential Indigenous population is the Dual System Estimator (DSE), sometimes referred to as 'dual survey estimators' or 'dual record systems', which can be used to benchmark our estimate of the Indigenous population within the New South Wales local court system.

The next section describes the broader issues for historical changes in the Indigenous population, which is followed by a detailed introduction to the ROD data and proposes that a statistical model be used to predict whether the unknowns are more like the Indigenous or non-Indigenous population. After presenting a summary of the results estimated from this model, the penultimate section benchmarks these results using a simple DSE that has been used to estimate populations for various groups in many countries (see Hunter & Dungey 2006). The final section provides some concluding remarks about the utility of the estimators used and points to future research directions that might prove useful for policy makers and researchers. While this paper attempts to build our confidence in the data so that we can feel confident in further analysis, we hope to illustrate a few basic issues that can and should be considered by anyone who collects and analyses data on Indigenous Australians.

THE BIG PICTURE FOR INDIGENOUS POPULATION CHANGES

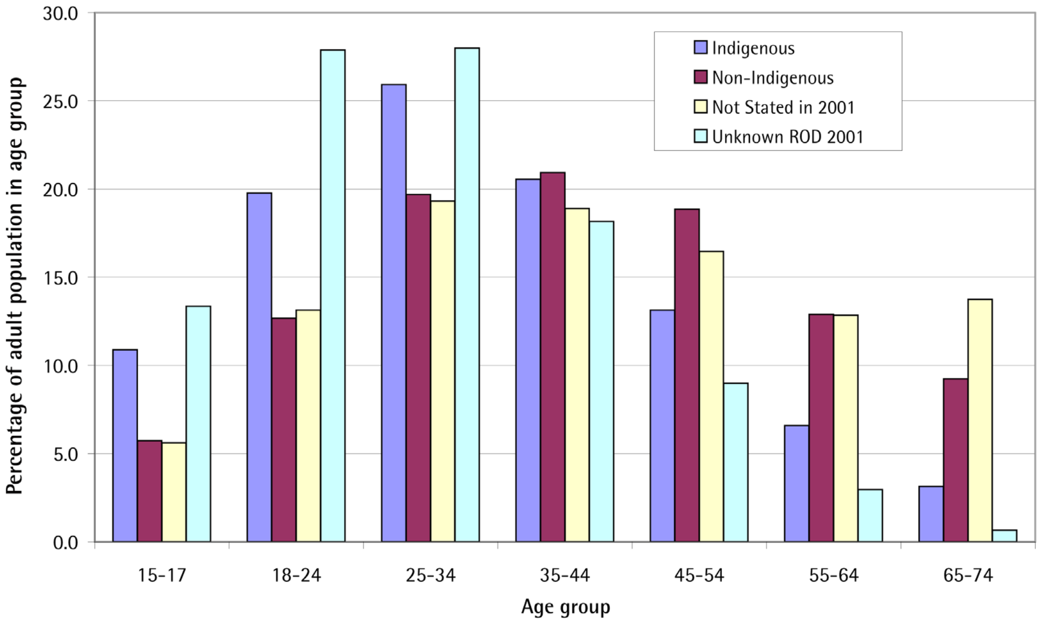
In general, population levels change over time according to the demographic balancing equation (Shyrock, Siegel & Associates 1976: 4):

$$\text{ERP}_{t+1}=\text{ERP}_t+\text{Births}_t-\text{Deaths}_t+\text{Net Migration}_t+\text{Census Procedure}_t+\text{E}_t \qquad (1)$$

where the ERP_t is the time-specific Estimated Residential Population and E_t is an error or residual term. The term ERP_t takes into account the tendency to miss some people when counting populations—that is, net undercount at a particular point of time. The recognition of this tendency does not deny the existence of double counting in some circumstances, but the reality is that many Indigenous people often do not identify or choose to identify as Indigenous. The balancing equation can be characterised as an accounting identity when one examines the total population because the residual term means that this equation is always true by definition. For sub-populations such as Indigenous and non-Indigenous population the changes are complicated by non-biological population growth (sometimes called an 'error of closure', which is largely embodied in the term E_t). This non-biological growth can include a component due to increased (or decreased) propensity to identify as Indigenous and another component due to inter-marriage between in various sub-populations—in particular, an increased rate of identification arising from the resulting progeny from such marriages. Another factor that effects non-biological growth is the change in both coverage of sub-populations (Guimond 1999) and the census editing procedures (Ross 1999). The census procedures in equation (1) are sometimes included in estimates of the 'error of closure' because it is difficult to get a precise measure of the effect of collection methodology on estimated populations.

The substantial 'error of closure' in Indigenous population levels means that trends in related outcomes are particularly difficult to interpret. This is because changes in an individual's Indigenous status over time lead to changes in the composition of the Indigenous population—which is difficult, if not impossible, to take into account before the population has identified itself at a particular point in time. Predicting future outcomes for the Indigenous Australian is fraught for the same reason in that one can never be sure about the extent of the non-biological growth. The main point is that it is difficult to evaluate long-running initiatives and changes in the policy regime because research can never be sure that cross sectional data measured over time relates to the same group of individuals. That is, longitudinal data are required in order to make sense of trends in Indigenous welfare and other outcomes.

Fig. 1. Demographic profile by Indigenous status from 2001 Census data and ROD



The census counts of Indigenous people, and the related issues of undercounts and the increasing propensity to identify as Indigenous, illustrate that many who do not self-identify as Indigenous in one census may eventually identify as Indigenous in a latter census.

Two questions that arise in the context of this paper are, firstly, who is likely to change their Indigenous status; and secondly, is the process of increasing propensity to identify as Indigenous beginning to wane?

With respect to the first question, it is worth noting that the level of non-response to the Indigenous status in the censuses is over twice the size of the actual number of people who identified as Indigenous directly. ABS (2007a) reports that 5.7 per cent of Australians did not respond to the question on Indigenous status in the 2006 Census (i.e. using usual residence counts).

With respect to the second question, the non-biological increase in the Indigenous population is never likely to be finalised, because intermarriage between Indigenous and non-Indigenous people means that the resulting children are likely to self-identify as Indigenous to some extent. There is a high and increasing rate of intermarriage, especially in urban areas.

Another general observation that can be made from the ABS (2007a) results is that the census questions for which respondents are less likely to respond are questions which relate only to part of the population, or for which respondents are uncertain about the appropriate response (e.g. residential status in non-private dwelling, unpaid domestic work and unpaid assistance to a person with a disability). The size of the non-response rates for Indigenous status question is not unduly large compared to other questions, but the non-response rate is particularly significant and has important implications given the number of non-responses relative to the number of people who identified as Indigenous.

This paper argues that those who do not indicate their Indigenous status in one year are the most likely part of the population to indicate they are Indigenous in future. Even if a relatively small proportion of these 'unknown' respondents change their status to Indigenous in future, the census based estimates of the Indigenous population will be measured with considerable error because there are so many 'unknowns'.

ATSI: Aboriginal and Torres Strait Islander

Fig. 1 shows the demographic profiles for Indigenous, non-Indigenous and unknown populations in the census and the unknown Indigenous status in the local court data used in ROD. People with unknown Aboriginal and Torres Strait Islander (ATSI) status in ROD are actually closer to the Indigenous profile than that for other Australians. That may be partially driven by the younger profile of the population involved in the criminal justice system. The importance of this figure is that it illustrates that it is still worth asking questions about the unknown categories and the remainder of this paper does that. Note that the following analysis refers to both ATSI status in local court data at a particular point in time, and the consolidated Indigenous status on ROD—the latter indicates whether a person identified as an Indigenous person at any appearance in the ROD.

Another observation that can be made in Fig. 1 is that the demographic profile of the not stated category in the census is not that different to the self-identified non-Indigenous population. Even if the unknown category is disproportionately Indigenous, the fact that Indigenous people are such a small minority of the Australian population means that the demographic profile for the not stated would in most circumstances be similar to the non-Indigenous profile.

RUMSFELD REVISITED, OR, KNOWING ABOUT THE 'UNKNOWN'

Donald Rumsfeld, the former US Secretary for Defence in the Ford and the George W. Bush Administrations, rather unfairly won the official 2003 Foot in Mouth Award for mangling the English language with his somewhat philosophical contribution to a press conference:

> Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know.

To our minds, Rumsfeld was just outlining the logical possibilities, or rather the combinations that one could enumerate the degree to which we think we understand (know) about what we think we 'know'. On the surface this may seem like Rumsfeld was engaging in some tautological discourse, but there is some subtlety underlying his somewhat awkward exposition.

In order to understand Rumsfeld we need to revisit some of the history of science. Hacking (1975) discusses the dualistic nature of probability where it can have both epistemological and aleatory dimensions. Epistemic probabilities concern our knowledge or what it is possible to understand, whereas aleatory probabilities have to do with the physical state of coins or mortal humans. In this way one can talk about the 'probability of various possibilities' (French Mathematician, Laplace, quoted in Hacking (1975): 133). If one is being charitable to Rumsfeld, one could also talk about the extent to which we know about 'knowns' and 'unknowns'. [2]

Identification of people as 'Indigenous' is not a physical state in any real sense, but is a reflexive exercise that may change over time. Therefore it is impossible for either an 'other' individual or society at large to 'know' whether people are Indigenous or not. Some people might suggest that genetic information gives a well-defined objective criteria for identifying indigeneity, but we believe that such is a misunderstanding of the nature of genetic expression and a rather deterministic notion of race that does not really stand up to scrutiny (Gould 1981).

The philosopher Willard Van Orman Quine is often quoted as writing that there is 'no entity without identity' (1981: 102). The quote resonates beyond the clever word play in that individuals need identity to exist as social beings, and hence our interactions with networks and communities are crucial determinants of our behaviour. We would take this further in that policy that ignores identity is likely to fail, as the people who are the subject of the policy will not take ownership of the issues underlying Indigenous disadvantage (Hunter 2007). This is an important point that underscores the importance of one of the major criticisms of the Northern Territory Emergency Response. However, in the context of this paper, the insight into the reflexive nature of identity provides a salutary warning that one should not excessively reify the nature of Indigenous identity. Notwithstanding, planning an effective Indigenous policy is not possible without some idea about the true Indigenous population.

DATA AND METHOD

OVERVIEW OF ROD

The ROD is not a simple data source that has been consistently collected and collated, but rather a compilation of several data sets all potentially constructed using different criteria. BOCSAR have done an invaluable job in combining the data so it is broadly comparable for major socio-demographic characteristics. Data from the Children's, Local and Higher Courts include a profile of sex, age, ATSI status and current location (measured at various levels—postcode, Local Government Area and Statistical Divisions) for the particular court appearance. These data sources also include information on offence and penalty, but such data was not used in this study as we were primarily interested in a person's individual characteristics and identity so that better estimates could be made of the Indigenous populations. Data from Youth Justice Conferences and custodial data from the New South Wales Department of Correctional Services (DCS) was also available but was not used for similar reasons.

DCS: Department of Correctional Services

The courts' ATSI status indicator is sourced from the police records (so when the matter goes to court, the ATSI status is filled in from the police file rather than the courts recording it separately). There is no audit between the court records and police records. The quality of the police use of this flag apparently increased after 1995, when a department level push began for an increase in the usage of ATSI status (pers. com. Weatherburn).

As noted above, BOCSAR constructed a consolidated measure of Indigenous status which identifies whether an individual ever had it recorded that they identified as ATSI at some stage in ROD data. In the remainder of this paper this will be referred to as the 'consolidated Indigenous identifier' whereas the raw police data will be referred to as the ATSI indicator. The DCS data also provided an additional (third) measure of Indigenous status that was independent of that recorded in police data.

The total number of unique individuals appearing each year in the ROD data increased gradually over the period examined in this paper from around 90,000 in 1994 to almost 110,000 in 2006.

It is important to understand the structure of the data used in this study. We did not follow the characteristics of individuals for every court appearance (at any point in time) in the study period as this would entail an enormous amount of data that would be difficult to manage. While BOCSAR had access to all the local court records in ROD we reduced the size and dimensionality of the data management exercise by focusing on the characteristics of people for their first appearance in a local court in all years between 1994 and 2006. Obviously, this means that our analysis did not capture all the information on ROD where there were multiple court appearances in a particular year. However, we would argue that this simplification is

Fig. 2. Number of years offenders appeared at least once in ROD

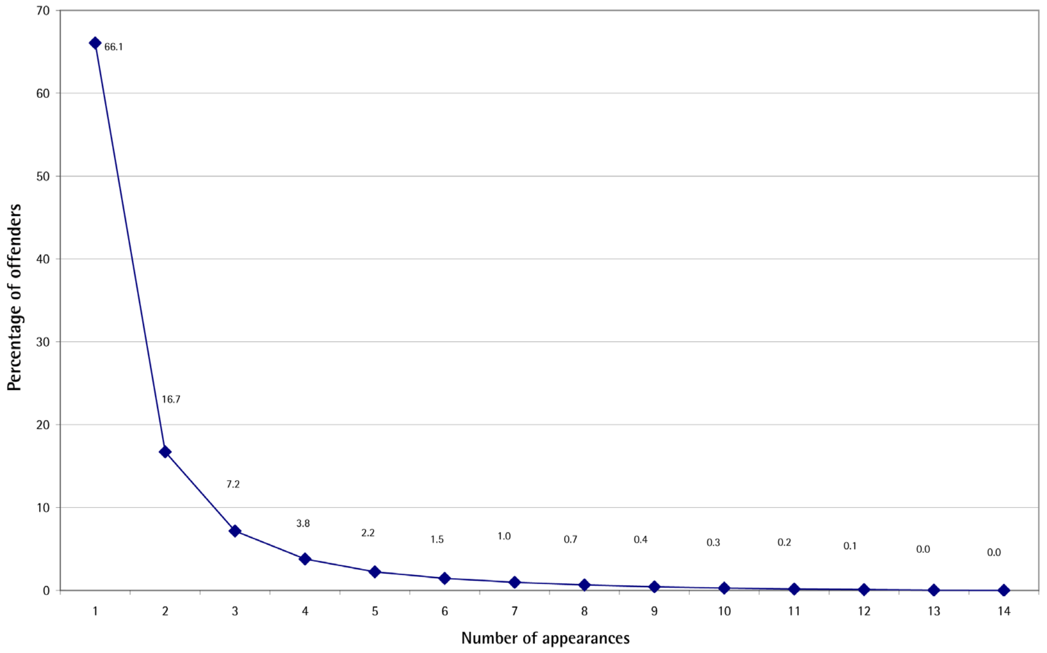


Fig. 3. Percentage of ROD individuals whose ATSI status is listed as unknown, 1994–2006

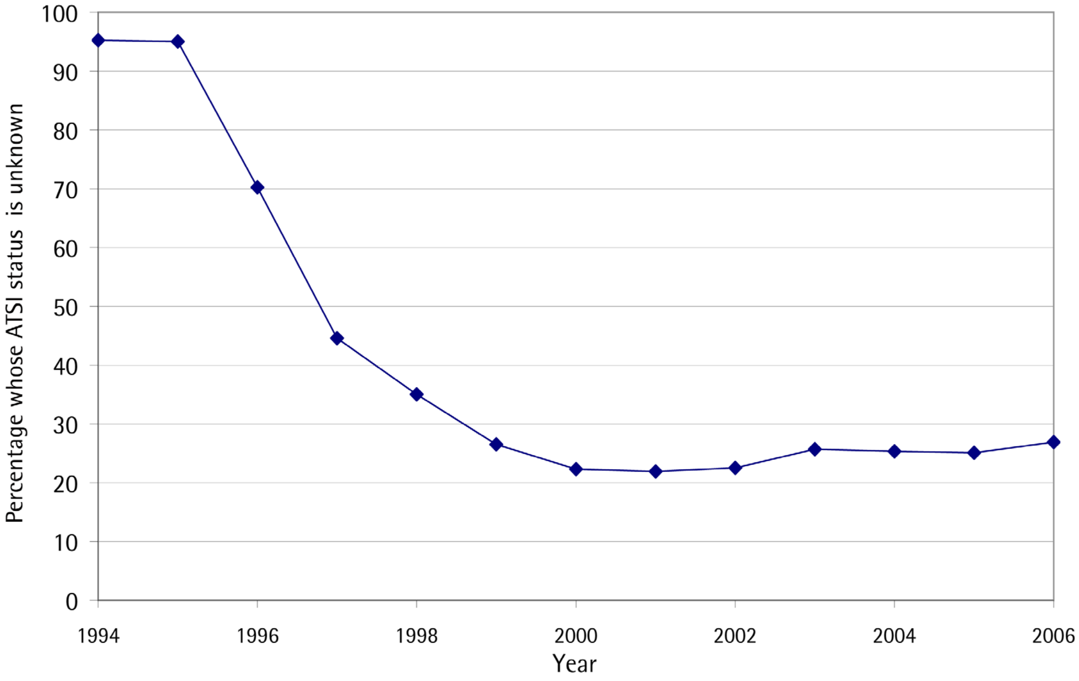
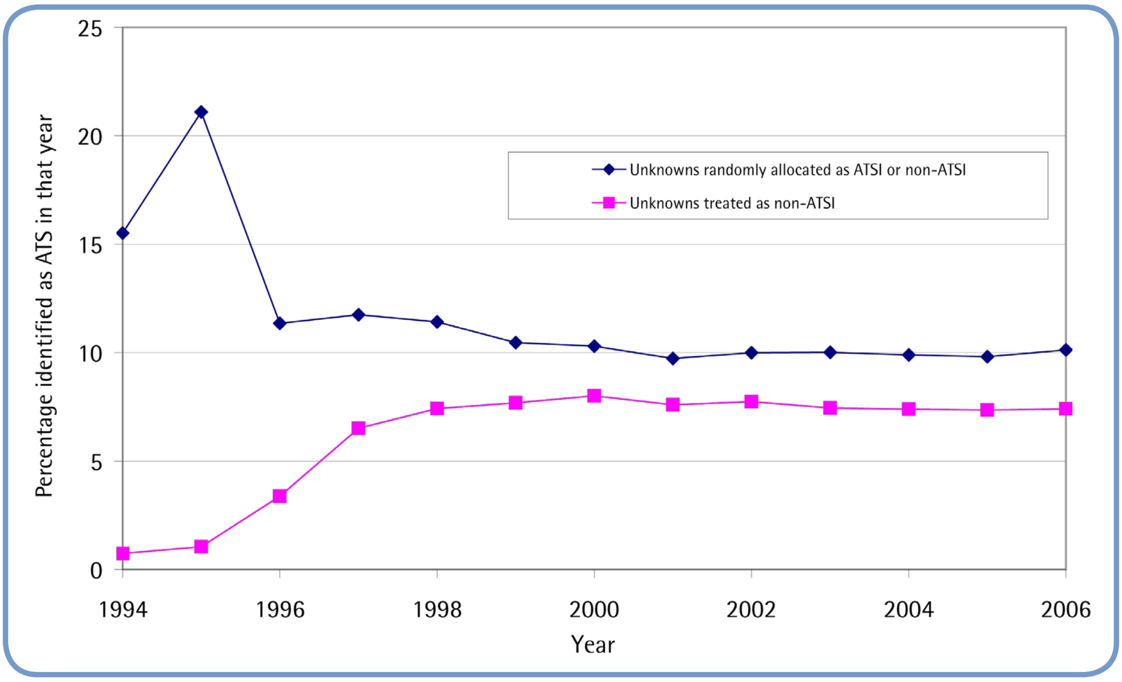


Fig. 4. Percentage of individuals in ROD identified as ATSI, using two methods of allocating persons with unknown ATSI status



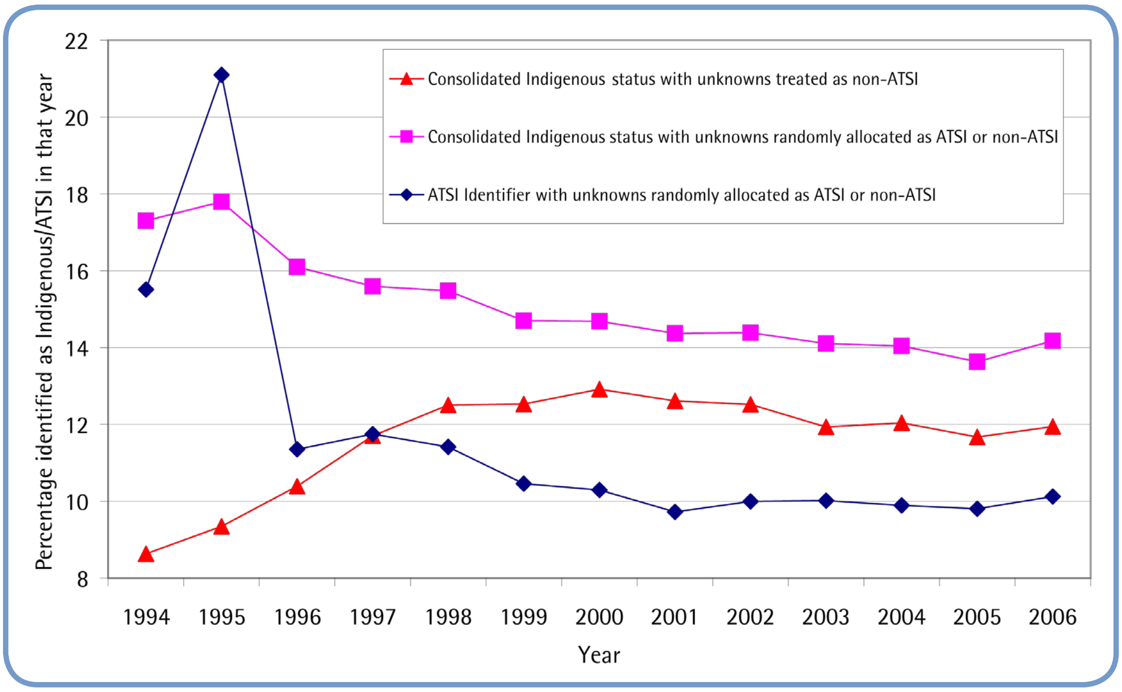
justifiable in that the information used in our study either does not change or is slow to change over time (demographic characteristics, Indigenous status and location of residence). By focusing on annual changes we can capture most of the variation in the data.

Of the 89,945 individuals recorded in our ROD data at some stage in 1994, less than 20,000 of these also appeared in local court in 1995. The number of these individuals in subsequent years declined over time but was generally over 10,000 people reappearing in any given year (the only exception was 2006, when the number of the original individuals appearing at least once was 9,579). While the rate of reoffence was quite high over the period examined, some of the original individuals in ROD in 1994 did not appear in later years. Of course, other people appeared for the first time in ROD in 1995 and later years. That is, the ROD data are intrinsically dynamic, and hence it is not easy to provide a summary overview.

Fig. 2 provides preliminary information on some of the dynamics within ROD. The mean number of years that a person appears in ROD can be defined as the total number of years people appear in court at least once divided by the unique persons in ROD. Given that there have been about 1.4 million appearances and 766,220 persons in ROD for the period examined the average person appears in slightly less than two separate years (1.78 years) during the period 1994–2006. Fig. 2 reports this information in a slightly different form in that two-thirds of people only appeared in ROD in one year—although these people appeared in court more than once by definition (as it is a Repeat Offender Database).

The fluctuations in the percentage of unknowns in ROD provide prima facie evidence of the data quality issues. Of course, if the percentage of unknowns in the administrative data is over 50 per cent there is literally more you do not know about the Indigenous population than you do know. There is no definitive rule whereby one can establish that administrative data are reliable, but if you take this time series in Fig. 3 as a starting point, then the percentage of unknowns decreases around 1997 or 1998 to around 30 per cent and stays stable at around 20-30 per cent thereafter. This observation accords with the above anecdotal evidence of a departmental push to use the ATSI indicator.

Fig. 5. Random allocation of unknown ATSI and Indigenous status



At the time of writing, a person for whom there is no information recorded on their ATSI status by the police is treated as non-Indigenous for the purposes of comparative statistics calculated from the ROD (Snowball & Weatherburn 2006: 6). The whole premise of this paper is that the unknown category should be interrogated to see if there is some unused information that would allow more robust and reliable comparative statistics to be calculated for both Indigenous and non-Indigenous populations. The simplest way to do this is to randomly allocate the unknown category between Indigenous and non-Indigenous groups, using a uniform statistical distribution, on the seemingly reasonable grounds that we do know whether such people are more likely to be one or the other. In other words, to assume the same percentage of Indigenous and non-Indigenous persons in both the known and unknown populations.

Fig. 4 reports the results of this calculation. In 1994, the randomly allocated results were 15 percentage points higher in estimates of the ATSI population than those which treated the unknowns as all being non-Indigenous. The size of this 'wedge' is due to the high levels of unknowns. Note that the difference in the estimated rate of ATSI status is greatly reduced by 1997 and seems to stabilise at just over 2 per cent of the population who appeared in local court at some stage during a particular year. There will always be a wedge between these estimates as long as there is some non-response to the questions about ATSI status. Given that we do not have any reason for expecting that the proportion of ATSI is going to vary much in the short–run, Fig. 4 would seem to add weight to our assertion that the data quality for ASTI status was not very good before 1997 and one should probably discount results generated for that period.

Fig. 5 compares the consolidated Indigenous identifier, both with and without the unknowns randomly allocated, against the ATSI indicator with unknowns allocated. The consolidated identifier usually implies a higher predicted Indigenous population within ROD than would be predicted using the percentage of ATSI estimated after the random allocation of unknowns. This is understandable, since people who indicate they are ATSI in latter (or previous) court appearances, but did not do so in the current year, are reclassified as Indigenous. However, once the unknown category is randomly allocated for the ATSI indicator and the consolidated Indigenous indicator, there is no reason why this should be the case. However, the only time when the allocated percentage of ATSI status is greater than the percentage with consolidated Indigenous status is in 1995. It is likely that the main reason for this is the exceptionally large number of people with unknown ATSI status in the local court data in that year.

In addition to providing the best guess of the true Indigenous population without any sophisticated statistical treatment, Fig. 4 confirms that the proportion of ROD with some Indigenous identity is relatively stable after 1997 or 1998. This is consistent with the above suggestion that the quality of the ROD data with respect to Indigenous status is credible and reasonably robust after 1997.

GEOGRAPHIC ISSUES

Postcode information is nominally available for ROD. However, there are several problems when using it in the context of Indigenous Australians. The main issue is that there are an insufficient number of Indigenous Australians in the average Australian postcode for many statistical purposes (see Hunter 1996). Neighbourhood analyses of postcode data are sometimes attempted for the total Australian population, but the lack of credible population estimates for the Indigenous Australians at this geographic level means that such analyses cannot conducted for Indigenous communities.[3]

The ABS draws boundaries and estimates the population distribution for other geographic levels of analysis. Local Government Areas (LGAs), which can be aggregated from Statistical Local Area (SLA) boundaries relatively easily, are also revised using the latest census data and the changes are recorded in the hierarchical Australian Standard Geographic Classification (ASGC). The LGA boundaries are relatively large and stable over time, compared with postcodes, and hence they are large enough to minimise measurement error.[4]

Another way to limit the error introduced by uncertainty over geographic boundaries is to use a robust indictor that captures broad spatial differences rather than measuring something that is specific to areas. The ABS Accessibility/Remoteness Index of Australia (ARIA) index is one such index. One indicator that is related to this ARIA index is the Levels of Relative Isolation (LORI) index that was used in Western Australian Aboriginal Child Health Survey (WAACHS) studies to characterise the accessibility of local Indigenous communities (Zubrick, Lawrence et al. 2004). Aggregations of areas classified by LORI also provide a meaningful indication of the accessibility of services in an area and hence the LORI index values for SLAs are aggregated to LGAs for use in what follows.

LGA: Local Government Area

SLA: Statistical Local Area

ASGL: Australian Standard Geographic Classification

ARIA: Accessibility/ Remoteness Index of Australia

LORI: Levels of Relative Isolation

WAACHS: Western Australian Aboriginal Child Health Survey

STATISTICAL MODEL OF UNKNOWNS

Some evidence is provided in Fig. 1. that individuals in ROD with unknown ATSI status may be closer to the Indigenous population in the census, at least in terms of their basic demographic age profile. However, this similarity may be a result of the distinctive nature of the local court data and hence it is probably necessary to estimate the basic demographic profiles for those for whom we have direct information regarding ATSI status. At a minimum one would expect that the distinctive demographic profiles of males and females also be taken into account when trying to estimate whether those of unknown ATSI status are closer to ATSI or other residents of New South Wales. However, as has been documented above, we have strong reasons to expect that the processes of identification are very different in more remote environments compared to the accessible areas of the state (see Ross 1999).

The following analysis uses a simple binary logistic framework to predict whether an individual with unknown status is either ATSI or non-ATSI using a series of socio-demographic and geographic characteristics for their first court appearance in a particular year.[5] Logistic regressions are often used where the dependent variable has two possible values, zero or one. For example, a person can either identify themselves as having either ATSI or non-ATSI status. To overcome the fact that this is a limited dependent variable, a logit transformation is used to ensure that the predicted probabilities lie between zero and one. The basic formulation of the binomial logistic regression model is

$$Logit\,P_i = \log\left(\frac{P}{(1-P)}\right)_i = bX_i - \qquad (1)$$

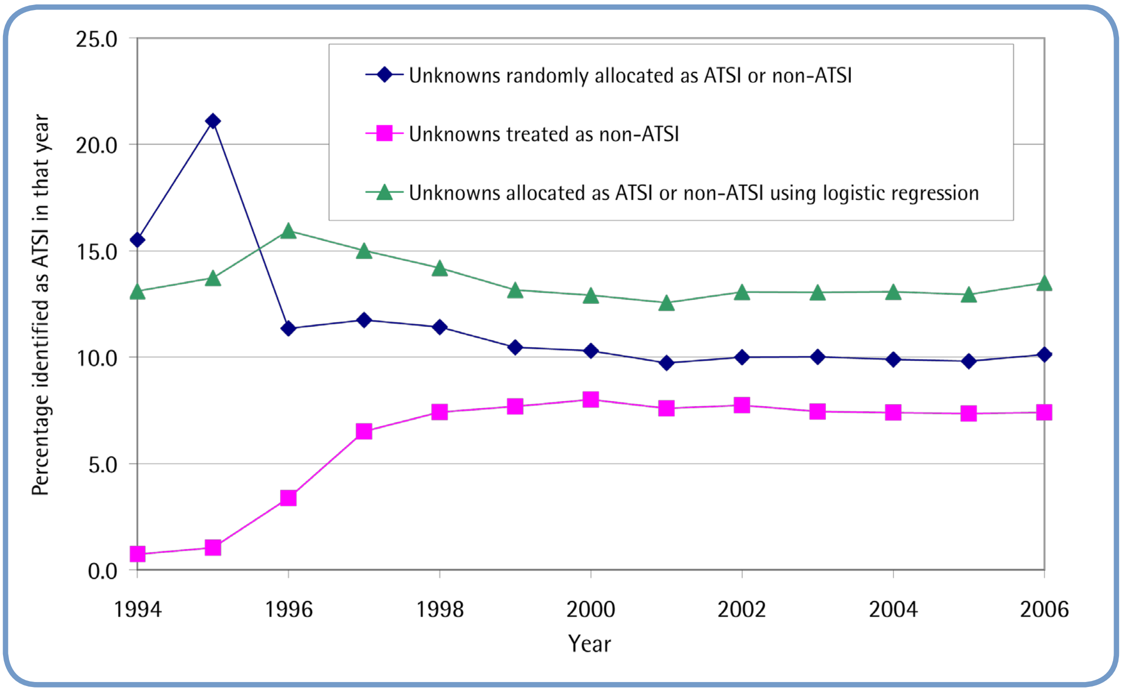
where *b* is a coefficient vector, the explanatory variables X_i and e_i are the error terms which approximate a normal distribution. (For fuller discussion, see Agresti 1984 and Hosmer & Lemeshow 1989). *Logit P*, also known as the log odds ratio, is the dependent variable in the logistic regression. The logistic regression models are estimated using maximum likelihood estimation techniques.

Often the coefficients of the binomial logistic model are interpreted using the log odds ratio. Hosmer and Lemeshow (1989) show that the log odds, or rather the natural log of the odds ratio, equals the individual coefficient of the respective variables.[6] When the explanatory variables are also binary, the coefficients in a logistic model must be interpreted as relative to a reference person defined by the omitted categories of the respective groups of explanatory variables. The reference person, or base case, in the following analysis is a female, aged 25-34 years living in a highly accessible SLA. Therefore, if we are interested in the effect of being male on the probability of being ATSI, then a negative coefficient implies that males are less likely to be identified as ATSI than females (i.e., the odds ratio is less than one).

While we are primarily interested in the characteristics of people more likely to be identified as ATSI, two regression models are estimated separately for all the years examined in this paper. The first regression model estimates whether the 'unknowns' are different from those who identified as either ATSI or non-ATSI in terms of their sex, age and remoteness/accessibility of residence at the time of the court appearance. If there are systematic differences, then there is a prima facie evidence for taking into account these characteristics when estimating the ATSI population of offenders to scope the significance of the issue. These regressions had significant coefficients for all the variables reported in Appendix Table A2. Males are slightly less likely to have unknown ATSI status compared to females. In most years those aged 55 or older were around twice as likely to have unknown ATSI status compared to the base age group of 25-34 year olds. Those in the youngest age group, 15-17 year olds, were around four or five times as likely to have unknown ATSI status until 2003, when they started becoming about as likely as the base age group to have unknown ATSI status. The remoteness indicator of geographic accessibility was significant for at least some categories in all the years examined. In general, at least from 2001 onwards, increasing the level of remoteness appears to be associated with an increased likelihood to have unknown ATSI status compared with those living in highly accessible areas. These results constitute a strong empirical rationale for providing some statistical underpinning to our assumptions about the unknown category, rather than relying solely on random allocation which does not take into account these socio-demographic differences.

In order to do this, we need a model to help us determine what we do know about the differences between ATSI and non-ATSI population in the ROD data given that a person has some valid ATSI status identified for their records. A second logistic regression model is therefore estimated to predict which factors are associated with an individual being more likely to be identified as ATSI as opposed to non-ATSI in our ROD data after excluding the unknowns. (See Appendix A3, which reports the odd-ratios and associated standard errors for a basic set of demographic and geographic characteristics). Males are about half as likely to be identified as ATSI as females. Also, in general the older a person is the less likely they are to identify as ATSI, with the youngest age group, 15-17 year olds, being around three to four times as likely to identify as ATSI compared to the base age group in all years of the analysis. Finally, increasing the level of remoteness is significantly (and substantially) associated with being more likely to identify as ATSI.

Fig. 6. Preliminary allocation of unknown ATSI status using regression estimates for 1994–2006



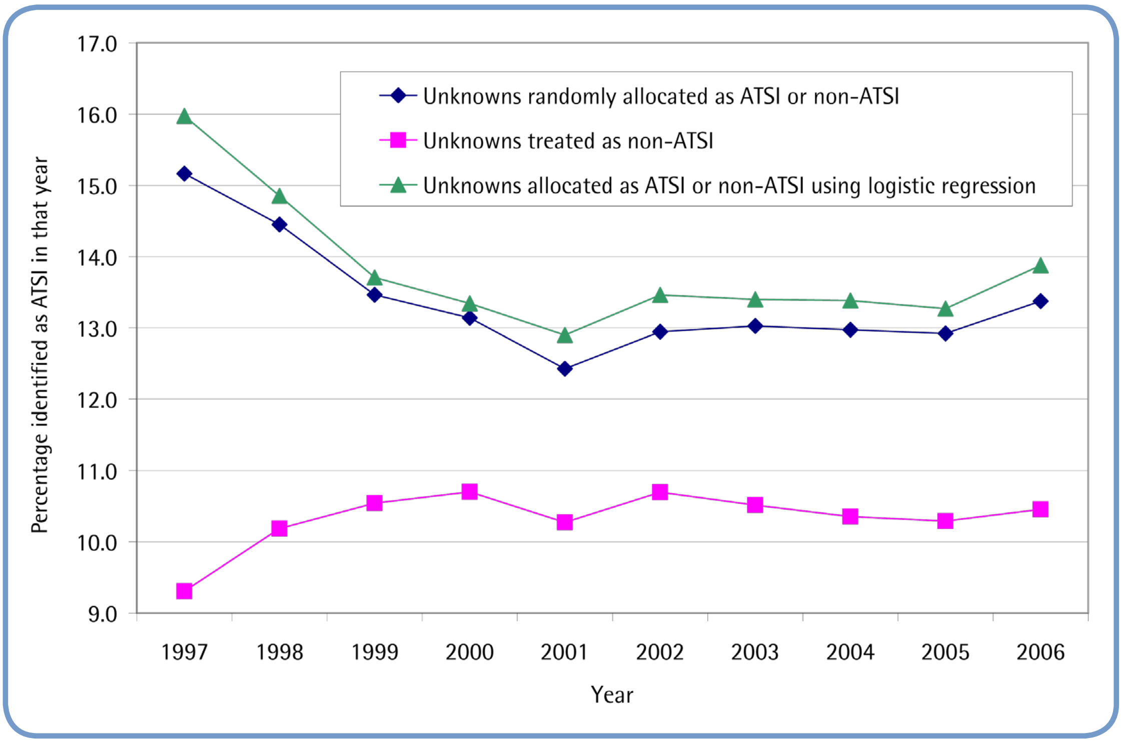
Note: The regression estimates used here are similar to those reported in Appendix A, which are confined to those individuals who appeared in the local court since 1997. These regression estimates use a more expansive sample that included the people who appeared in ROD since 1994 and for which there was the necessary information required to reclassify unknowns as either ATSI or non-ATSI.

The next step in the analysis is to provide an 'out-of-sample' prediction of the proportion of ATSI for those with unknown ATSI status in the respective years. That is, we then ask whether the people with unknown ATSI status are more like the ATSI or the non-ATSI populations, and classify the person as such for the purposes of estimating the true population.

Before reporting the results, we reflect briefly on the implications of the fact that the data used is highly grouped and hence there are relatively few cells from which to impute ATSI status. As a result the predicted probabilities for such cells will be classified as either ATSI or non-ATSI (depending on whether the probability of being ATSI is greater than 0.5). There is nothing particularly wrong with this procedure except that the small number of cells means that it will lead to less reliable estimates than would otherwise be the case. Consequently, we use an alternative method, whereby we first estimate the probability of being ATSI for each cell (i.e., conditioned on relevant demographic and geographic characteristics), and then multiply by the number of those with unknown status in that cell. While this should provide a reliable and robust estimate, it is rather difficult to estimate the standard error for the overall estimates. This does not matter excessively, since this exercise is designed to illustrate the potential importance of the issue. However, given the large number of unknowns and the associated low standard errors for the estimates, it is anticipated that this estimator provides a highly accurate estimate of the number of unknowns re-classified as ATSI.

Fig. 6 reports some preliminary estimates of the effect of re-classifying unknowns for ROD for the period 1994-2006. The important thing to note about this preliminary analysis is that it includes the period about which we have concerns regarding data quality in ROD, that is 1994-1996. Even so the percentage of ATSI

Fig. 7. Estimated ATSI identification in ROD (%), 1997–2006



Note: The regression estimates used in this figure are reported in Appendix A. The sample only included the people who appeared in ROD since 1997 and for which there was the necessary information required to reclassify unknowns as either ATSI or non-ATSI.

after logistic reclassification is quite stable for this period and not necessarily totally inconsistent with the estimates for the entire period examined. Overall, a naïve interpretation of Fig. 6 would seem to indicate that there are many unknowns that are re-classified as ATSI using the regression model rather than the random allocation technique—indeed, the percentage of ROD that were identified as ATSI appeared to be between two and five percentage point higher when the logistic regression method was used. This inference would not be valid because the random allocation was estimated for the whole ROD, whereas the regression estimates focus on the sub-sample for which we have no missing observations for the demographic and geographic characteristics.

Given that Fig. 6 is potentially misleading in the inferences about the extent to which unknowns are reclassified as ATSI, Fig. 7 reports the levels of reclassification using the logistic and random allocation techniques after restricting the focus solely to those individuals for which we have complete information on sex, age and geography. Since we have good reasons to expect that the pre-1997 data are of poor quality, Fig. 7 uses only the data included in the regression estimates provided in Appendix Table A3.

One of the main points is that, when one compares Figs 4–6 with Fig. 7, the estimated percentage of ATSI is higher in the latter after the random allocation procedure is performed on the unknown ATSI category. The main reason for this is that the latter was confined to those for which all the relevant demographic and geographic data was available. For such data the proportion of people who identified as ATSI increased by around three percentage points, irrespective of the allocation of the unknown categories.

We surmise from the small size of the difference between the random allocation and logistic allocation of ATSI status that the diligent recording of demographic and geographic data for the local courts is associated with more diligence in recording ATSI status. Hence the auditing of records to improve the overall quality of demographic data will also improve the reliability of evidence for Indigenous over-representation in the criminal justice system.

As indicated above, the ROD data reports a consolidated Indigenous identifier that takes into account previous identification patterns in New South Wales local court data. The above analysis can also be conducted for the distribution of this indicator. Note that one would expect this indicator to deliver a higher level of Indigenous identification, and hence it is less likely to involve an undercount. In most circumstances this measure of Indigenous identification would be expected to be closer to the true Indigenous population within ROD—although, a priori, one cannot reject the hypothesis that changes in Indigenous status are not occurring in an arbitrary fashion that is unrelated to individual's true identity. Fig. 7 illustrated that the extent of reclassification when using regression estimates is only slightly higher than that done when unknowns are randomly assigned (using uniform distribution). Given that there are fewer unknowns to assign when the consolidated Indigenous identifier is used, it is reasonably certain that there will not be much difference between the randomly allocated and regression adjusted estimates of the proportion of Indigenous people in ROD. Another reason not to estimate the regression adjustments for the consolidated Indigenous identifier is that these are cross-sectional regressions, and the ROD ATSI status variable uses information across time. Therefore any regression adjustments are likely to be correlated over time, which would certainly induce less reliability in the resulting estimates and may induce some bias with more recent offenders having had less time for their consolidated Indigenous identifier to be updated or 'consolidated'. Notwithstanding, as the above indicates, the estimates based on the random allocation of unknowns for the consolidated Indigenous identifier are likely to provide a close approximation of our best guess for the true Indigenous population within New South Wales local courts.

THE DUAL SYSTEM ESTIMATOR METHODOLOGY

The following uses the DSE to validate the above estimates of the Indigenous population within the ROD data. The simplest DSE is a two-sample model. The first sample identifies certain individuals who are returned to the population after the survey is complete, while the second sample provides an independent measure of the population. Using the numbers of individuals in both samples and the numbers identified in just one sample, it is possible to estimate the number not captured in either sample, thus providing an estimate of the total population size. The assumptions required for such an estimate to be valid are that:

1. there is no change to the population during the investigation (i.e. the population is closed)
2. individuals can be matched from one sample to the next
3. the chance of being in each sample is uncorrelated for each individual, and
4. the two samples are independent.

Sekar and Deming (1949) were the first to adapt the method for human populations when they used the method to estimate birth and death rates, and the extent of their registration in 1949, with hospital data from India. There is also a substantial literature, going back to the 1940s, dealing with the application of the two-sample method to census data (Fienberg 1992). By taking another sample in addition to the census, the method can be used for estimating undercount by the census (Hogan 1993).

Table 1. A two outcome example of DSE methodology

	Response A		
	Yes	No	Total
Response B			
Yes	x_{11}	x_{12}	$x_{11} + x_{12}$
No	x_{21}	x_{22}	$x_{21} + x_{22}$
Total	$x_{11} + x_{21}$	$x_{12} + x_{22}$	$x_{11}+x_{12}+ x_{21}+ x_{22}$

In terms of the validity of assumptions for estimating the potential numbers of Indigenous Australians, it is necessary to confine our attention to closed populations. Even populations with high mobility, such as people in remote Indigenous communities, may be considered 'closed' so long as the PES or follow-up survey takes place shortly after the initial survey or census (Paradies et al. 2000).

With respect to assumption (2), matching will depend on the quality of the records and the uniqueness of respondents' names. BOCSAR has given a detailed assurance that all due care is taken to match the local court data for individuals by spending considerable time and resources constructing a unique person identifier.

Another of the assumptions required for DSEs to be valid is the homogeneity of the population (assumption 3 above). That is, all the members in the population should have the same chance of being sampled in the follow-up survey. However, this assumption is unlikely to be violated as it is not a choice for most offenders. Both the police and the DCS records ATSI status as accurately as possible, although there is obviously room for doubt about the certainty of the categories, as is evidenced by the existence of the unknown category for both 'surveys'.

The question of independence is discussed by Sekar and Deming (1949) in some detail (also see Marks et al. 1974). As indicated above, the assumption of independence seems to be valid in this case.

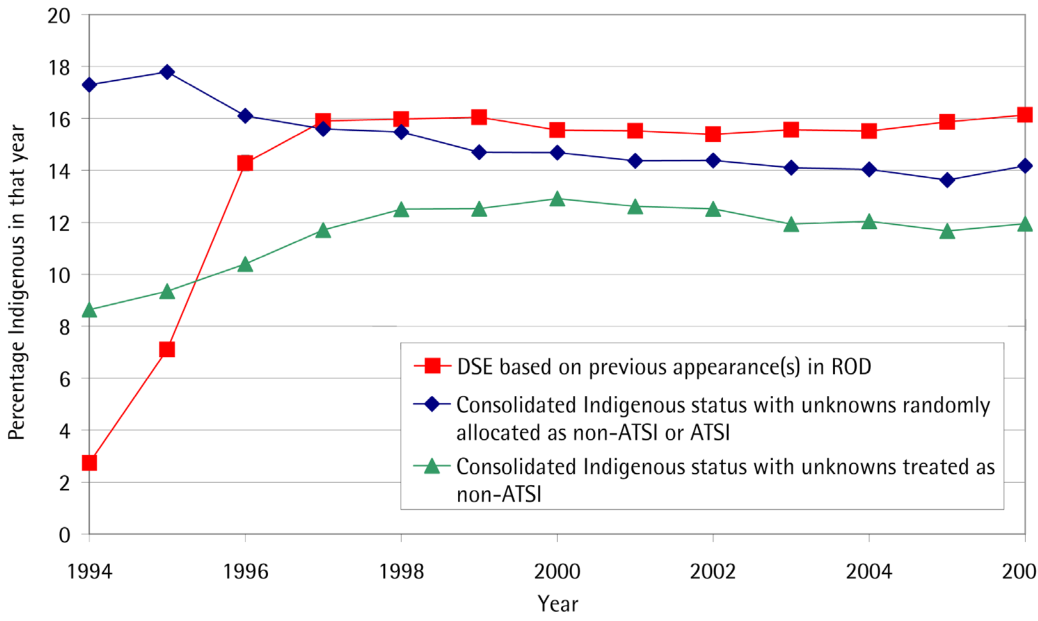
Another issue that can cause problems for the DSE methodology is that of coverage. If there are individuals who are not sampled in both sets, this results in potential upwards bias of the estimates (Shyrock, Siegel & Associates 1976). For census data and administrative data, which in principle covers the whole population, this source of error should be relatively small.

The key to the DSE method is an ability to match individual records, on some different criteria (i.e. different to the one of immediate interest), and then check the observation of interest for consistency. In a two-outcome situation, such as a yes/no question, four potential outcomes occur, as illustrated in Table 1. First, the record can be 'yes' on both the initial and second surveys, designated by the cell x_{11}. Second, the record can be 'yes' on the first and 'no' on the second, designated by the cell x_{12}. Third, the record can be 'no' on the first and 'yes' on the second, denoted by cell x_{21}, and finally the record can be 'no' on both surveys, given by x_{22}. This method cannot, of course, pick up information that has been incorrectly recorded on both surveys (e.g. respondents answering 'yes' on both surveys when the true observation was 'no').

Using the Sekar–Deming (1949) formula, the revised population estimate is:

$$N = x_{11} + x_{12} + x_{21} + x_{12}x_{21}/x_{11} \qquad (2)$$

Fig. 8. DSE estimates of ATSI population consistent with consolidated Indigenous indicator (after allocation)



Note: The 95 % confidence intervals for the DSE are approximately equal to the size of the boxes used to de-mark the entries for particular year. They are actually reported in this figure but the confidence intervals are encompassed by the year markers.

If Table 1 refers to the response to a question about Indigenous status, then only x_{22} people always deny they are Indigenous. Consequently, Hunter (1998) referred to the potential Indigenous population as being equal to $x_{11} + x_{12} + x_{21}$. The consolidated Indigenous identifier on ROD is closely related to this 'potential Indigenous population'. The main difference is that the ROD estimate is potentially based on repeated appearances in local court and hence takes into account more than two 'surveys'. However, some of the x_{22} people may also admit to being Indigenous in other circumstances (i.e., if other similar independent surveys were conducted repeatedly). The 4th term on the right-hand side of equation 2 is the number expected to identify as Indigenous at least once if all surveys are 'independent' (in statistical terms). The variance of can be estimated using the standard binomial approach (see Sekar & Deming 1949).

The DSE of the percentage of ROD that are Indigenous are reported in Fig. 8. These are remarkably stable and we would argue that it provides the most accurate estimate of the true population of Indigenous offenders in ROD. It is probably not a coincidence that the DSE based on any previous appearance is almost identical to the estimates based on the consolidated Indigenous identifier after random allocation in 1997. Before that year the high level of unknowns and poor data quality of local court records, especially with respect to Indigenous status, lead to unreliable estimates for both the consolidated Indigenous estimates and the DSEs. The DSE is particularly affected in those years because it is driven by the exceptionally small number of people specifically identifying as ATSI in the earlier years. In the context of this DSE, the process of predicting whether unknowns could be assigned to ATSI or non-ATSI is more complex than the cross-section-based estimates reported above, and requires a more sophisticated technique. Hence it is left for another paper.

Even if one discounts the reliability of the estimates for the pre-1997 period, observant readers will note that there is a gradual decline in the percentage of ROD estimated to be Indigenous after the consolidated Indigenous identifier is allocated randomly. We suspect that this is due to the fact that the consolidated Indigenous identifier is less likely to assign an Indigenous status if there have been fewer years over which people could change their status. In a sense, the data are more affected by right-censoring when individuals only enter the local court system for the first time in the years immediately leading up to 2006. The DSE seems to be less affected by this distortion because it is only defined for people who have appeared at least twice (i.e., over several years).

REFLECTIONS ON KNOWING SOMETHING ABOUT THE UNKNOWNS

This paper argues that it is important to understand the processes that determine who is identified as or chooses to identify as Indigenous. If nothing else the size of Indigenous involvement in the criminal justice system will be severely underestimated if no attempt is made to establish or estimate the true identity of the large number of people with unknown ATSI status within the criminal justice system. There is an auditing process for police records and local court data, but this cannot entirely remove the uncertainty and hence the relative over-representation of Indigenous and non-Indigenous populations needs to involve some statistical procedure to estimate the true Indigenous population of offenders.

The main conclusion of this paper is that Indigenous disadvantage will be severely understated if administrative data are not corrected to account for those who may at some later stage identify as Indigenous or whose Indigenous status is unrecorded or unknown. The secondary message is that data quality issues not only decrease the reliability of resulting estimates for Indigenous and other Australians, but also result in the potential for systematic biases which could affect conclusions about the size of Indigenous disadvantage and the ability of policy makers to 'close the gap'. These observations are particularly important for the administrative data collections reported in the OID Framework, and the policies that arise from such statistical reportage (e.g. SCRGSP 2007).

These conclusions are underscored by the significance of geographic factors in the processes that determine Indigenous status—both in the model of whether ATSI status was unknown and in the model of whether an offender was identified as ATSI given that their status was 'known'. The geographic distributions of ATSI status will be systematically biased with respect to the incidence of unknown status and the incidence of ATSI (given that ATSI status is 'known') as both are significantly correlated with the local accessibility of the local geographic area. That is, inferences from the unadjusted ROD data about relative crime rates for Indigenous and non-Indigenous people will not be valid.

The reliability of measures of Indigenous disadvantage is further complicated by the need to estimate local ERPs for the Indigenous and other population for use in the calculations of rates of offence in the respective populations. The calculation of accurate ERPs for the Indigenous population is itself hotly debated in the academic and administrative literature (Taylor 1997), but the failure to use ERP for the local Indigenous population will result in distorted pictures of Indigenous involvement in the criminal justice system. Given that the Indigenous ERPs usually experience higher undercounts than is evident for general ERPs, Indigenous rates are highly likely to be overstated by more than they are for the rest of the population (ABS 2007b). While some might argue that one should not worry too much at an aggregate level as Indigenous disadvantage is such a manifest problem, such distortions may disproportionately affect certain regions, and hence administrative data needs to be as accurate and reliable as possible. However, a more accurate estimate of Indigenous offender populations could be achieved by alternative methods, including the allocation of unknowns or by using a DSE methodology. Such estimates should be combined with local ERPs estimates for Indigenous and other Australians to ensure that policy to address relative offence rates is only based on valid empirical evidence.

There are other methods for estimating offender populations which could be considered (Collins & Wilson 1990). However, such methods are valid for estimating the unobserved Indigenous and non-Indigenous rates by estimating the Indigenous and other Australians who do not appear in the court or are identified in police records (using count data models). While such methods are invaluable for estimating consistent offender rates in the relevant populations, and obviates the need to generate consistent and comparable ERPs, the above analysis is justified solely as an exercise in validating the quality of ROD Indigenous identifier given that a person has been through the local court system.

One possible next step for research is to use a regression-based model of DSE to use the information about Indigenous unknown status in that framework. That is, as the above DSE estimates are not based on a reclassification of unknowns for ROD and the DCS, researchers could consider using an iterative model of such reclassification using the overall information in ROD. It would be preferable to separately model all the changes in identification between known and unknown status. As this involves a more complicated statistical model, with a 3x3 matrix of possible Indigenous status, it is beyond the scope of this paper, and hence is will have to be the subject of a subsequent paper.

While it is intrinsically difficult to 'know the unknown', the cost of not attempting to understand the consequences of the category 'unknown' is likely to be misleading and potentially very costly in terms of the failure to properly allocate resources for designing an effective Indigenous policy.

NOTES

1. The Australian PES is an interviewer-based survey conducted three weeks after census night which allows comparison of the responses in the census and the PES to identify whether they have changed. Here we use a matched sample of those who responded to both the census and the PES. It is also possible that the PES may pick up some uncounted population from the census—as both samples are drawn from the population as a whole. Information is collected to determine whether persons have been missed or double counted in the census and whether dwellings were missed. The PES collects personal information on Indigenous origin, age, sex, marital status and birthplace. Note that there are several differences between the census and PES collections. For example, the census question on Indigenous status is based on self-identification whereas the PES involves an interviewer. In addition there were slight differences in the wording of the question. More importantly, the PES question is asked of the entire household whereas the census is asked of each person individually.
2. One comment that we cannot resist making is that one possible combination of words that Rumsfeld omits is that he does not consider the unknown 'knowns'. However, that omission is understandable in that it may bring up images of the collective unconscious and universal archetypes. Given that we also do not want to discuss the analytical psychology tradition (founded by Carl Jung), we only pursue the categories used by Rumsfeld.
3. Postcodes are a difficult spatial unit to work with as it is difficult to ensure clarity and continuity of boundaries through time. Postcodes change a fair bit over time and really only exist as a hard geographic entity at the time of the various censuses. Postcodes are a important administrative unit with boundaries that arguably exist most cogently in the minds of Australia Post staff. There has been a concerted effort to regularise the boundaries of residential postcodes in recent years, most notably with the greater use of ABS postal areas, but 'administrative' postcodes that are dedicated to post office boxes (and what are sometimes called large volume receivers) complicate matters because the spatial dimensions of such postcodes are not always clear—hence, assumptions about the physical location of such residences usually have to be made.
4. In contrast to postal areas, LGA boundaries are relatively slow to change over time. LGA-level ROD data are estimated by BOCSAR.
5. It could be argued that the analysis of the unknown category in ROD should investigate all the information available on the unknown category that is associated with Indigenous status. For example, the incidence of being recorded as unknown Indigenous status seems highly associated with offence-type and the propensity to offend. The inclusion of such factors may ultimately enhance the specification, but it is also possible that the processes for recording Indigenous status are correlated to the processes for recording offence type (especially driving offences) and the statistical processes that drive to propensity to offend. In Econometrics this would raise the possibility of simultaneous equation bias (Greene 2000: 710). In an attempt to avoid such issues, this paper uses an extremely parsimonious specification that should be accurate on average. A more sophisticated analysis might be able to discount the possibility of simultaneity bias and hence could confidently use additional information on offence type and propensity to offend.
6. See Hosmer and Lemeshow (1989) for details of the interpretation of these ratios.

APPENDIX A. DESCRIPTIVE STATISTICS AND REGRESSION ANALYSIS FOR PREDICTING ABORIGINAL/TORRES STRAIT ISLANDER STATUS IN A PARTICULAR YEAR FOR ROD, 1997–2006

Table A1. Descriptive statistics: Number of ATSI, non-ASTI and unknown (Part I)

	1997			1998			1999		
Variable	Non-ATSI	ATSI	Unknown	Non-ATSI	ATSI	Unknown	Non-ATSI	ATSI	Unknown
Sex									
Female [1]	2,907	964	2,586	3,993	1,194	2,394	5,252	1,505	1,973
Male	21,818	3,455	15,751	27,898	4,193	13,217	34,263	4,643	10,684
Age									
15 to 17 year olds	953	602	3,670	1,363	754	3,072	1,537	763	2,205
18 to 24 year olds	9,355	1,500	5,574	11,758	1,730	4,620	14,381	2,003	3,706
25 to 34 year olds [1]	8,492	1,556	4,860	10,666	1,910	4,079	13,521	2,153	3,334
35 to 44 year olds	4,181	613	2,769	5,778	796	2,432	7,072	953	2,096
45 to 54 year olds	1,350	134	1,094	1,810	169	1,037	2,344	239	949
55 to 64 year olds	333	9	294	439	22	305	551	32	300
65 to 74 year olds	61	5	76	77	6	66	109	5	67
Remoteness (ARIA)									
Highly Accessible [(a)]	15,113	1,100	9,589	18,687	1,359	8,529	23,091	1,538	7,250
Accessible	9,045	2,376	8,092	12,622	3,051	6,658	15,751	3,596	5,053
Moderately Accessible	517	702	571	521	730	333	636	797	250
Remote	50	241	85	61	247	91	37	217	104
Very Remote	0	0	0	0	0	0	0	0	0

Notes: These statistics relate to the ROD data that has complete information on sex, age and geography and hence were used in the regression analysis in Appendix Tables A2 and A3.

1. Base categories are females aged 25 to 34 years and living in areas clarified as highly accessible. Insignificant results in bold & standard errors in brackets.

Table A2. Descriptive statistics: Number of ATSI, non-ASTI and unknown (Part II)

	2000			2001			2002		
Variable	Non-ATSI	ATSI	Unknown	Non-ATSI	ATSI	Unknown	Non-ATSI	ATSI	Unknown
Sex									
Female [1]	5,280	1,481	1,705	5,729	1,491	1,724	5,773	1,565	1,819
Male	34,698	4,567	8,783	37,185	4,599	8,560	37,240	4,831	8,569
Age									
15 to 17 year olds	1,402	733	1,361	1,190	636	1,373	990	572	1,290
18 to 24 year olds	14,023	1,946	2,980	14,594	1,909	2,868	14,037	1,921	2,841
25 to 34 year olds [1]	13,740	2,106	3,033	14,903	2,179	2,879	14,971	2,281	2,755
35 to 44 year olds	7,617	1,008	1,901	8,413	1,074	1,867	8,812	1,214	2,049
45 to 54 year olds	2,518	219	886	3,012	255	925	3,214	355	1,020
55 to 64 year olds	546	34	270	671	31	304	816	51	362
65 to 74 year olds	132	2	57	131	6	68	173	2	71
Remoteness (ARIA)									
Highly Accessible [(a)]	23,791	1,560	5,809	26,143	1,659	5,232	26,414	1,792	5,289
Accessible	15,543	3,558	4,377	16,136	3,538	4,664	16,040	3,709	4,709
Moderately Accessible	592	704	234	582	713	296	506	656	296
Remote	52	226	68	53	180	92	53	239	94
Very Remote	0	0	0	0	0	0	0	0	0

Notes: These statistics relate to the ROD data that has complete information on sex, age and geography and hence were used in the regression analysis in Appendix Tables A2 and A3.

1. Base categories are females aged 25 to 34 years and living in areas clarified as highly accessible. Insignificant results in bold & standard errors in brackets.

Table A3. Descriptive statistics: Number of ATSI, non-ASTI and unknown (Part III)

	2003			2004			2005			2006		
Variable	Non-ATSI	ATSI	Unknown	Non-ATSI	ATSI	Unknown	Non-ATSI	ATSI	Unknown	Non-ATSI	ATSI	Unknown
Sex												
Female [1]	5,655	1,482	2,011	5,664	1,521	2,083	5,739	1,576	2,011	4,945	1,421	2,027
Male	36,902	4,893	9,685	36,246	4,727	10,111	35,959	4,612	10,231	31,746	4,244	9,800
Age												
15 to 17 year olds	1,224	644	756	1,199	582	756	1,238	647	429	782	437	231
18 to 24 year olds	13,175	1,947	3,094	12,597	1,939	3,283	11,703	1,806	3,127	9,153	1,596	2,742
25 to 34 year olds [1]	14,718	2,118	3,458	14,493	2,053	3,705	14,460	2,018	3,844	13,080	1,905	3,961
35 to 44 year olds	8,871	1,284	2,496	8,907	1,275	2,598	9,274	1,270	2,704	8,736	1,260	2,765
45 to 54 year olds	3,392	327	1,309	3,563	343	1,270	3,743	379	1,448	3,719	399	1,437
55 to 64 year olds	995	49	475	961	49	479	1,086	65	565	1,021	62	562
65 to 74 year olds	182	6	108	190	7	103	194	3	125	200	6	129
Remoteness (ARIA)												
Highly Accessible [(a)]	26,488	1,891	5,704	26,817	1,856	5,777	26,420	1,939	5,733	23,224	1,798	5,361
Accessible	15,625	3,644	5,548	14,661	3,564	5,989	14,688	3,360	5,892	12,895	3,070	5,742
Moderately Accessible	397	621	364	373	588	362	518	648	521	489	615	623
Remote	47	219	80	59	240	66	55	235	86	65	179	93
Very Remote	0	0	0	0	0	0	17	6	10	18	3	8

Notes: These statistics relate to the ROD data that has complete information on sex, age and geography and hence were used in the regression analysis in Appendix Tables A2 and A3.

1. Base categories are females aged 25 to 34 years and living in areas clarified as highly accessible. Insignificant results in bold & standard errors in brackets.

Table A4. Odds ratios of unknown (vs known) ATSI status, 1997–2006

VARIABLE	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Male	0.91	0.87	0.92	0.87	0.85	0.81	0.83	0.86	0.94	0.88
	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)
15 to 17 year olds[1]	4.90	4.51	4.56	3.36	4.43	5.14	1.92	1.83	**0.94**	0.69
	(0.17)	(0.15)	(0.16)	(0.13)	(0.18)	(0.22)	(0.09)	(0.09)	(0.05)	(0.05)
18 to 24 year olds	1.06	1.06	1.07	**0.98**	**1.03**	1.12	**0.99**	**1.00**	**0.98**	0.95
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
35 to 44 year olds	1.18	1.14	1.23	1.15	1.16	1.27	1.19	1.13	1.10	**1.04**
	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
45 to 54 year olds	1.53	1.62	1.74	1.70	1.69	1.80	1.73	1.46	1.52	1.33
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.06)	(0.05)	(0.05)	(0.05)
55 to 64 year olds	1.78	2.05	2.43	2.46	2.60	2.69	2.30	2.19	2.18	2.03
	(0.15)	(0.16)	(0.18)	(0.19)	(0.19)	(0.18)	(0.13)	(0.13)	(0.12)	(0.11)
65 to 74 year olds	2.39	2.46	2.81	2.24	2.96	2.58	2.83	2.37	2.79	2.49
	(0.40)	(0.41)	(0.44)	(0.36)	(0.44)	(0.37)	(0.35)	(0.29)	(0.33)	(0.29)
Accessible[1]	1.20	**0.97**	0.86	**0.98**	1.24	1.25	1.43	1.63	1.63	1.70
	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
Moderately Accessible	0.77	0.59	0.57	0.76	1.18	1.31	1.76	1.85	2.27	2.67
	(0.04)	(0.04)	(0.04)	(0.06)	(0.08)	(0.09)	(0.11)	(0.12)	(0.12)	(0.14)
Remote	0.47	0.64	1.32	**0.94**	1.95	1.52	1.45	**1.08**	1.47	1.85
	(0.04)	(0.04)	(0.04)	(0.06)	(0.08)	(0.09)	(0.11)	(0.12)	(0.12)	(0.14)
Very Remote	-	-	-	-	-	-	-	-	2.19	**1.69**
	-	-	-	-	-	-	-	-	(0.83)	**(0.71)**
Number of observations	47,481	52,889	58,320	56,514	59,288	59,797	60,628	60,352	60,128	54,183
Pseudo R2	0.04	0.04	0.04	0.02	0.03	0.03	0.02	0.02	0.02	0.02

Note: 1. Base categories are females aged 25 to 34 years and living in areas clarified as highly accessible. Insignificant results in bold and standard errors in brackets.

Table A5. Odds ratios of ATSI vs non-ATSI status, 1997–2006

VARIABLE	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Male	0.48	0.52	0.48	0.46	0.47	0.47	0.51	0.48	0.48	0.47
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
15 to 17 year olds[1]	3.10	2.72	2.90	3.18	3.54	3.63	3.35	3.09	3.69	3.97
	(0.20)	(0.15)	(0.16)	(0.18)	(0.21)	(0.23)	(0.20)	(0.19)	(0.21)	(0.27)
18 to 24 year olds	0.89	0.82	0.87	0.88	0.88	0.90	**1.01**	1.09	1.09	1.19
	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.05)
35 to 44 year olds	0.78	0.76	0.82	0.82	0.81	0.88	**0.98**	**1.00**	**0.97**	**0.96**
	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
45 to 54 year olds	0.56	0.53	0.64	0.55	0.56	0.70	0.67	0.69	0.72	0.76
	(0.06)	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)
55 to 64 year olds	0.15	0.28	0.37	0.41	0.30	0.42	0.34	0.36	0.42	0.44
	(0.05)	(0.06)	(0.07)	(0.08)	(0.06)	(0.06)	(0.05)	(0.06)	(0.06)	(0.06)
65 to 74 year olds	**0.47**	0.42	0.29	0.10	0.28	0.08	0.21	0.29	0.13	0.20
	(0.23)	(0.18)	(0.14)	(0.07)	(0.12)	(0.06)	(0.09)	(0.11)	(0.08)	(0.09)
Accessible[1]	3.46	3.18	3.34	3.43	3.44	3.37	3.19	3.39	3.06	3.05
	(0.14)	(0.11)	(0.11)	(0.11)	(0.11)	(0.10)	(0.10)	(0.10)	(0.09)	(0.10)
Moderately Accessible	17.95	18.63	18.92	18.33	19.80	19.34	21.28	22.24	16.44	16.16
	(1.21)	(1.22)	(1.15)	(1.16)	(1.25)	(1.26)	(1.49)	(1.60)	(1.07)	(1.08)
Remote	66.09	54.27	86.99	63.59	52.03	63.95	63.41	57.51	57.66	34.36
	(10.62)	(8.02)	(15.81)	(10.09)	(8.39)	(9.97)	(10.47)	(8.59)	(8.93)	(5.15)
Very Remote	-	-	-	-	-	-	-	-	3.73	**2.14**
	-	-	-	-	-	-	-	-	(1.87)	(1.35)
Number of observations	29,144	37,278	45,663	46,026	49,004	49,409	48,932	45,158	47,886	42,356
Pseudo R2	0.16	0.14	0.13	0.13	0.13	0.13	0.12	0.13	0.12	0.12

Note: 1. Base categories are females aged 25 to 34 years and living in areas clarified as highly accessible. Insignificant results in bold and standard errors in brackets.

REFERENCES

Australian Bureau of Statistics (ABS) 2007a. *Non-response Rates, Australia 2006 Usual residence and Place of Enumeration*, cat. no. 2914.0.55.001, ABS, Canberra.

—— 2007b. *Population distribution, Aboriginal and Torres Strait Islander Australians 2006*, cat. no. 4705.0, ABS, Canberra.

Agresti, A. 1984. *Analysis of Ordinal Categorical Data*, Wiley, New York.

Altman, J.C., Biddle, N., and Hunter, B.H. 2008. 'How realistic are the prospects for 'Closing the gaps' in socioeconomic outcomes for Indigenous Australians?', *CAEPR Discussion Paper No. 287*, CAEPR, ANU, Canberra, available at <**http://www.anu.edu.au/caepr/discussion.php**>.

Collins, M.F. and Wilson, R.M. 1990. 'Automobile theft: Estimating the size of the criminal population', *Journal of Quantitative Criminology*, 6 (4): 395–409.

Fienberg, S.E. 1992. 'Bibliography on capture-recapture modeling with application to census undercount adjustment', *Survey Methodology*, 18 (1): 143–54.

Gould, S.J. 1981. *The Mismeasure of Man*, Penguin, London.

Greene, W.H. 2000. *Econometric Analysis*, Prentice Hall, New Jersey.

Guimond, É. 1999. *Ethnic mobility and the demographic growth of Canada's aboriginal populations from 1986 to 1996*, cat. no. 91-209-XPE, Statistics Canada, Ottawa.

Hacking, I. 1975. *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*, Cambridge University Press, London.

Hogan, H. 1993. 'The 1990 post-enumeration survey: Operations and results', *Journal of the American Statistical Association*, 88: 1047–60.

Hosmer, D. and Lemeshow, S. 1989. *Applied Logistic Regression*, Wiley, New York.

Hunter, B.H. 1996. 'Indigenous Australians and the socioeconomic status of urban neighbourhoods', *CAEPR Discussion Paper No. 106*, CAEPR, ANU, Canberra, available at <**http://www.anu.edu.au/caepr/discussion.php**>.

—— 1998. 'Assessing the utility of 1996 Census data on Indigenous Australians', *CAEPR Discussion Paper No. 154*, CAEPR, ANU, Canberra, available at <**http://www.anu.edu.au/caepr/discussion.php**>.

—— 2007. 'Cumulative Causation and the Productivity Commission's Framework for Overcoming Indigenous Disadvantage', *Australian Journal of Labour Economics*, 10 (3): 185-202.

—— and Dungey, M.H. 2006. 'Creating a sense of 'CLOSURE': Providing confidence intervals on some recent estimates of indigenous populations', *Canadian Studies in Population*, 33 (1): 1–23.

Marks, E.S., Seltzer, W., Krótki, K.J. 1974. *Population Growth Estimation: A Handbook of Vital Statistics Measurement*, The Population Council, New York.

Paradies, Y., Huppatz, S., Warnsey J. and Barnes, T. 2000. Population and globalisation: Australia in the 21st century, Paper delivered at the 10th Biennial Conference of the Australian Population Association, Melbourne.

Quine, W.V. 1981. *Theories and Things*, Harvard University Press, Cambridge.

Ross, K. 1999. *Occasional Paper: Population Issues, Indigenous Australians*, cat. no. 4708.0, ABS, Canberra.

Sekar, C. and Deming, E.W. 1949. 'On a method of estimating birth and death rates and extent of registration', *Journal of the American Statistical Association*, 44 (1): 101–15.

Shyrock, H.S., Siegel, J.S., and Associates. 1976. *The Methods and Materials of Demography*, Academic Press, London.

Snowball, L. and Weatherburn, D. 2006. 'Indigenous over-representation in prison: The role of offender characteristics', *Crime and Justice Bulletin*, 99 (September): 1–20.

Steering Committee for the Review of Government Service Provision (SCRGSP) 2003. *Overcoming Indigenous Disadvantage: Key Indicators 2003 Report*, Productivity Commission, Melbourne.

—— 2005. *Overcoming Indigenous Disadvantage: Key Indicators 2005 Report*, Productivity Commission, Melbourne.

—— 2007. *Overcoming Indigenous Disadvantage: Key Indicators 2007 Report*, Productivity Commission, Melbourne.

Taylor, J. 1997. 'The contemporary demography of indigenous Australians', *Journal of the Australian Population Association*, 14(1): 77–114.

Zubrick, S.R., Lawrence, D.M., et al. 2004. The Western Australian Aboriginal Child Health Survey: The Health of Aboriginal Children and Young People, Telethon Institute for Child Health Research, Perth.