The relationship between settlement type and undercount in the South African census of 2011

Risenga Maluleke & Amanda van Eeden

Peer reviewed and revised

Abstract

Holding a population and housing census is both a momentous undertaking and a costly challenge for any government. It requires vast planning and organising, but the results are vital for constructive planning. The aim of a census is to collect, process and disseminate detailed statistics on population size, composition and distribution at small-area level. As an undercount could affect the trust and use of census data, a major challenge in any census is dealing with the effects of an undercount. This article investigates some of the causes of the undercount in the South African National census of 2011, and how these causes vary across different geographic areas. The aim of the investigation is to determine the relationship between settlement type and the undercount in the 2011-census with the intention of laying the groundwork for lower undercounts in future censuses and survey implementation strategies. The key objectives of this article are to determine whether geographic location affects the census undercount and to understand how results based on geographic location are distributed spatially across the country.

DIE VERWANTSKAP TUSSEN NEDERSETTINGSTIPOLOGIE EN DIE ONDERTELLING VAN DIE SUID-AFRIKAANSE SENSUS VAN 2011

Die hou van 'n bevolking- en behuisingsensus is beide 'n belangrike onderneming en 'n duur uitdaging vir enige regering. Dit vereis groot beplanning en organisering, maar die resultate is noodsaaklik vir konstruktiewe beplanning. Die doel van 'n sensus is om gedetailleerde statistiek oor die bevolkinggrootte, -samestelling en -verspreiding oor 'n klein gebied te versamel, te verwerk en te versprei. 'n Groot uitdaging in enige sensus is die hantering van die gevolge van 'n ondertelling omdat 'n ondertelling die vertroue in en die gebruik van sensusdata kan beïnvloed. Hierdie artikel ondersoek sommige van die oorsake van die ondertelling in die Suid-Afrikaanse Nasionale sensus van 2011, en hoe hierdie oorsake varieer tussen verskillende geografiese gebiede. Die doel van die ondersoek is om die verhouding tussen die nedersettingstipe en die ondertelling in die 2011-sensus vas te stel met die bedoeling om 'n grondslag te bepaal vir laer ondertellings in toekomstige sensusse en opname implementeringstrategieë. Die belangrikste doelwitte van hierdie artikel is om te bepaal of geografiese ligging die sensusondertelling beïnvloed, en om te verstaan hoe die resultate gebaseer op geografiese ligging ruimtelik versprei word oor die hele land.

KAMAHANO MMOHO EA MEFUTA EA BOLULO LE HO SE BALE KA HO LEKANA PALONG EA SECHABA, AFRIKA BORWA, SELEMONG SA 2011

Ho tsamaisa palo ea sechaba le ea mehaho eo batho ba phelang ho oona, ke ntho e kholohadi ebile e ja chelete e ngata ho 'muso o mong le o mong. Ke ntho e hlokang ditukiso tse tebileng fela qetellong sephetho sa teng ke se mererong e hlophisitsoeng. Sepheo sa palo ea sechaba ke ho akaralletsa, ho sebebetsana le ho tlaleha dipalopalo tse phethahetseng tsa batho ka hara naha, le hore na batho ba bokellane ka mokhoa oo joang libakeng tse nyane ka hara naha. Ka ha ho se bale ka ho lekana ho ka ama ho sebelisa dipalopalo tsa sechaba ka bots'epehi, bothata ba ho ho se bale ka ho lekana bo tlameha ho qojoa mme se batleloe pheko ea mathata ao e se bakang ha e se e etsahetse. Serapa sena se shebisisa ditlamorao tsa ho se bale ka ho lekana palong ea sechaba, Afrika Borwa, selemong sa 2011, le hore na ditlamorao tsena di fapana joang di karolong ka ho fapana tsa naha. Sepheo sa chebisiso ena e ne e le ho bona kamahano ea mefuta ea bolulo le ho se bale ka ho lekana palong ea sechaba ea selemo sa 2011, ka morero oa ho beha dinomoro tsa ho se bale ka ho lekana tlase le mekhoa ea ho fumana mohlodi oa pokello hore palo tsa sechaba tsa bokamoso di tle di phethahale hantle. Dipheo tse bohlokoa tsa serapa sena e ne e le (i) ho bona hore na ho se bale ka ho lekana ho amme libaka tsa naha ka ho fapana; le (ii) ho utloisisa hore na diphetho tsa naha di ka hara naha joang.

1. INTRODUCTION

"Census taking is a momentous and challenging undertaking. It brings on board diverse political, economic and social demands in terms of the resources that have to be deployed to collect the requisite data" (Masiteng & Kekovole, 2006: 2). A population and housing census aims to collect, process and disseminate detailed statistics on population size, composition and distribution at small-area level. Data is used for countless purposes by a myriad of end-users but, in particular, by government for purposes of policy formulation, planning and monitoring of development outcomes.

In South Africa, the first census under the democratic order was conducted in 1996. However, a census does not enumerate everyone it intended to and this then leads to an undercount (Razi, 1999: 1106; Carter, 2009: 4)[1]. An undercount refers to the difference between the published census population estimates and the real numbers of people in the population. This is partly due to coverage non-response, whereby some people and dwellings are missed by the census operation in its entirety: this is not directly observable during census processing (Carter, 2009: 5). By contrast, underenumeration refers to missing responses, which should be identifiable within census processing and amenable to correction by estimation. Statistical agencies use different terms when estimating the number of people counted in the census. The terms

1 Many factors can be the cause of an undercount in a census undertaking. Statistics South Africa used the factors that contributed to the undercount in the 1996 and 2001 censuses to implement strategies to deal with the undercount in the 2011 census (for detail on these factors and the strategies, see Statistics South Africa 2009).

Mr Risenga Maluleke, Statistics South Africa, De Bruyn Park Building, 170 Thabo Sehume Street, Pretoria, 0002, South Africa. Email: <Risenga@statssa.gov.za>

Mrs Amanda van Eeden, Centre for Regional and Urban Innovation and Statistical Exploration (CRUISE), Department of Geography and Environmental Studies, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa. Phone: 02108089104, email: <avaneeden@sun.ac.za>

are commonly used, but not always well defined (Carter, 2009: 4). The terms 'undercount', 'non-response' and 'underenumeration' will now be briefly mentioned.

The 1996-census yielded an undercount of 10% (Statistics South Africa, 2009: 7)[2]. A subsequent census was conducted by Statistics South Africa (Stats SA) in 2001 and recorded an undercount of 17.9%, while a large-scale sample survey conducted in 2007 returned an undercount of approximately 18%. The latest census conducted in 2011 recorded an undercount of 14.6%. The successful censuses of 1996 and 2001 had considerable challenges in the new democratic dispensation that had dawned in 1994 (Christopher, 2009: 106-107). The results of these censuses added value to the understanding of the society, its economy, transformation and the effects of governance and state delivery. A high level of confidence in census data is crucial to enable effective planning, programming, monitoring and development by both public and private organisations. However, a significant undercount, such as the ones measured, dampens the credibility of the census as one that has 'counted everyone'; high undercounts affect the trust and use of census data.

Many factors such as gender, race, settlement type (urban, rural and remote areas), income and level of education can contribute to an undercount. Such factors can cause an undercount to vary across space and among variables (UN, 2001: 3; Koch & Cebula, 2004: 576; Carter, 2009: 10). All these variables, with the exception of gender and race, are sensitive to change.

Stats SA has just completed the processing of the 2011-census results – a process that included a strategy to deal with the undercount. Settlements are classified according to the degree of planning for a community organisation, the identification of tasks that provide the reasons for the existence for such a settlement, and the nature of the social organisation (Fried, 1963: 93).

In the 2006-census, Australia recorded that people in rural and remote areas contributed to undercount, whereas the United Kingdom and the United States, in their censuses of 2001 and 2000, respectively, observed that inner-city areas and highly concentrated urban poor areas contributed to undercount (Carter, 2009: 7-9). In South Africa, some of the settlement types categorised as 'hard-to-count' include homeless people, people living in high-walled areas (gated communities and residential complexes), informal settlements, remote rural communities and commercial farms (Stats SA, 2009: 30).

South Africa, and Stats SA in particular, has not conducted a detailed formal study to understand the undercount and its root causes. By scrutinising results from the Post-Enumeration Survey of 2001 (PES), the Community Survey Publicity Follow-up Survey, Census 2001 Debriefing Report as well as discussions within Stats SA (Stats SA, 2012), a number of factors have been identified as contributing causes of the undercount. For various reasons, the following issues seemed to be the most problematic: enumeration methods in certain areas; interpretation of census questions; unclear concepts and definitions; errors in the data processing and lack in information technology; limited publicity; uneven recruitment and training of enumerators; uncertainty about the physical characteristics to be enumerated in terms of population and settlement type; methods and timing of the PES, as well as mapping, demarcation and listing of dwellings (Carter, 2009: 3; Martin, 2010: 2754; Carr-Hill, 2012: online).

Post-Enumerator surveys (PES) is a household survey run immediately following a census. The PES estimates both the overcount and the undercount to provide government with a net undercount figure (Hogan, 1993; Brown, Eaton, Freedman, Klein, Olshen, Wacher, Wells & Ylvisaker, 1999: 354). Estimates of undercount are also used to derive adjustment factors for estimates of the resident population for the census year and evaluate the effectiveness of census collection procedures so that improvements can be made to future censuses (Isaki & Schultz, 1986; Hogan, 1993; Bell, 1993: 110).

This article aims to establish the relationship between the geographic location (urban and rural communities) and a high census undercount as one of the reasons potentially contributing to such an undercount. These results could inform the strategy used in subsequent censuses as well as in the deployment of resources and publicity efforts in areas that have a high undercount. Two issues are investigated, namely whether settlement types affect the census undercount over the three census years, and whether there is a relationship between different types of settlement and their effect on census undercount. The latter was done using the Chi-square Automatic Interaction Detection (CHAID) technique.

2. BACKGROUND TO CENSUS UNDERCOUNT

When a census is performed, it cannot enumerate everyone it intended to count. This then leads to an undercount (Carter, 2009: 7; Martin, 2010: 2755). In other words, undercount in a census occurs when a percentage of the people in the country are not counted at the time of the census (Razi, 1999: 1106).

This could be attributed to various reasons. For instance, people worldwide have inherited taboos about counting, especially the counting of people and livestock. Both people and livestock are about preservation of heritage and wealth; the majority of cultures believe that counting will attract the attention of bad omens and evil spirits and that, as a consequence of counting, some of them will disappear or perish (Frazer, 1918). In 1908, Congo state officials were bewildered at the resistance of the local population to be counted for tax purposes; the officials were of the opinion that the people were resisting tax, whereas they were resisting to be counted (Frazer, 1918). Similar experiences were noted among the Masai, Akamba, Sania Galla, and Kikuyu of Kenya as well as the San and the Koi of southern Africa (Frazer, 1918). Non-committal behaviour and refusal to participate in censuses is often rooted in the culture of the society, thus leading to undercount.

In England, a bill to introduce a census in 1753 was passed by the House of Commons and rejected by the House of Lords who opined that a census would provide information on military matters to England's enemies. The English Parliament was opposed to the concept of the census and believed that it could lead to misfortune. As such, England did not conduct a census until 1801 (Wrigley & Schofield, 1989: 3).

2 According to the UN Principles and Recommendations for Population and Housing Censuses, a net undercount occurs when omissions exceed the sum of duplications and erroneous inclusions.

The USA first acknowledged an undercount in their census of 1940, despite the fact that President George Washington and Secretary of State Thomas Jefferson had raised this matter 150 years earlier. The USA's 1940-census noted that 5.4% of the population had not been counted (Razi, 1999: 1106). The subsequent censuses of 1950, 1960, 1970 and 1980 lowered the undercount considerably. The USA census of 2000 noted an undercount of between 0.96% and 1.4%; thus, it had dropped from the undercount of 1.6% recorded in 1990 (US Census Bureau, 2000). The post-enumerated survey (PES) in the USA is a political issue as it influences the adjustments to the data collected during the census and thus translates into budgetary benefits or loss. States, cities and interest groups, who perceive that they would gain by it, advocate the PES, whereas those who believed that they would lose are opposed to it (Choldin, 1994).

Research has indicated that, although many countries have similar challenges, different groups of individuals within those countries can be identified as being undercounted (Martin, 2010: 2755; Simon, 2012: 1385; Carr-Hill, 2012). The 2006 New Zealand census reported an undercount of 2%. Men, especially of Asian and Pacific descent, were more undercounted than women, and there was a reported undercount of dwellers of high-rise apartments, especially in Auckland (Statistics New Zealand, 2007). Men of Asian and Pacific descent are more likely to have entered the country illegally, whereas access to high-rise apartments was a challenge. The United Kingdom census of 2001 noted an undercount of 6.1%, although a considerable number of strategies were put in place to deal with this matter. The UK identified 'hard-to-count' groups and adopted measures to deal with each group. Other key contributors to the undercount were inner-city dwellers as well as deficiencies in the address register (UK Statistics Commission, 2004).

Table 1 shows the undercount levels of countries where censuses were conducted in line with the United Nations guidelines during the 2000 Round of Population and Housing Census that started in 1995 and was concluded in 2004. Despite different methodologies employed by different countries to determine the undercount, global patterns indicate an undercount of less than 8% in those censuses. The majority of the countries take 2 to 3 years to calculate their undercount rate for a census; therefore, the latest figures for the majority of these countries are not available.

Table 1: Undercount levels internationally

Country	Year of census	National undercount
Australia	2001	2.2%
Canada	2001	3%
Mauritius	2000	2.5%
Mozambique	1997	5.1%
Nepal	2001	5.3%
New Zealand	2001	2.2%
Seychelles	2002	2.4%
Tanzania	2002	7.0%
USA	2000	1.2%
South Africa	2001	17.6%

Sources: Australian Bureau of Statistics, 2011; South African Government News Agency, 2012

3. MEASURING MODELS TO DETERMINE CENSUS UNDERCOUNT

Undercount can be estimated by means of two main methods, namely the Demographic Analysis (DA) and the Dual System Estimation (DSE). While considerable efforts have been made to combine the DA and the DSE methods to improve census estimates of the population, both methods have presented challenges (Isaki & Schultz, 1986: 173). The DA provides estimates exclusively at national levels of aggregation, whereas the DSE method requires modelling assumptions that are difficult to assess based on fit to the data (Elliot & Little, 2005: 381). A brief discussion of the DA and the DSE methods will be followed by a discussion of a more recent method used by Bycrofft (2006) to calculate undercount. This will be followed by a discussion of the Chi-square Automatic Interaction Detection (CHAID) technique used in this investigation.

3.1 The Demographic Analysis Method (DA)

The DA uses data on births, deaths and migration to estimate the undercount (Ahmed, Gupta, Robinson & Woodrow, 1993). It moves from a premise that:

Population = Births - Deaths + Immigration - Emigration

In this model, the population at time (t) is the population at time t-1 suitably aged for the duration of the period, plus births minus deaths and plus inward migration minus outward migration.

Therefore, P (t) = P(t-1) + Births – Deaths + In-migration – Out-migration

Starting with a base census population for t=0 and subsequently adjusting it for births, deaths and migration over a long period depends on the quality of the starting point data and requires an extremely high-quality registration system. Any inadequacies in the census base starting point will flow directly through to subsequent population estimates (Bycroft, 2006: 34). Any inadequacies in the records of external migration may well accumulate over the lifetime of the population and, like inaccuracies in birth and death registrations, may become apparent in old age either as negative populations or as unrealistic survival.

3.2 The Dual System Estimation Method (DSE)

Agencies that are responsible for undertaking censuses attempt to adjust census counts in order to reduce differential undercount, using a DSE method based on a capture-recapture method. Being counted in the census is considered *capture*; being counted in the PES is considered *recapture*. DSE involves the following:

- Taking a random sample of Enumeration Areas (EAs);
- Trying to enumerate the residents of those EAs after census day, in the PES;
- Trying to match PES records to census records, on the basis of

data that are often incomplete or erroneous;

- Estimating the undercount within demographic groups called post-strata by comparing capture-recapture estimates with census counts.

The method of estimating net undercount from PES data is known as DSE, as it uses two surveys, namely the census and its related PES, to estimate net undercount. The DSE works on the premise of an assumption of independence between the probability of response to census and the probability of response to PES. This assumption of independence is necessary to ensure that neither taints the other, that is, to ensure unbiased estimates of census undercount (Bell, 1993: 107; Elliot & Little, 2005: 381, Brown *et al*., 1999). A lack of independence between census and PES leads to 'correlation bias' in the DSE estimate, thereby leading to a lack of trust in both the surveys. While every effort is made to ensure the integrity of the PES and separation from the census, operational independence cannot be fully guaranteed. Bell (1993) describes two mechanisms that would lead to correlation bias from failure of a general independence assumption:

- Causal independence: the act of being included in the census makes one more likely or less likely to be included in the PES; or
- Heterogeneity: census and PES inclusion probabilities vary over persons within the DSE post-strata.

Two samples are critical in estimating an undercount, namely the *P-sample* and the *E-sample* (Fay, Passel & Robinson, 1988: 45-62; Hogan, 1993: 1050). Countries such as the US, the UK and Australia have applied these samples. Generally, the two samples exist for the following reasons:

- The *P-sample* provides an estimate for census omissions of persons and housing units. It is based on listed housing units and persons within them, independent of the census enumeration.
- The *E-sample* provides an estimate of erroneous enumerations, enumerations of persons who should not have been counted or who should have been counted elsewhere. It is based on census enumerations, partially overlapping with the *P-sample*.
- A specific choice of the DSE to combine results from the *P-sample*, the *E-sample*, and the census itself into an estimate of the true population.

3.3 New methods of measuring undercount

According to Bycroft (2006), a new method can be used to measure net undercount at national level without relying on a PES or on data on annual migration. The method has two distinctive features. The first is that its estimates of numbers of migrants from a particular country are based on figures of persons born within such a country and recorded as residents during censuses of other countries at a given point in time. The second feature is the Bayesian approach which espouses that each of the uncertain elements in the calculations is given an *a priori* error distribution and that a number of empirical constraints are imposed on the gender-age profile of percentage net undercount.

An undercount can be accounted for by a particular group of people within a society or by variables of questions that are asked in the census questionnaire. For purposes of population, demography and social studies, undercount is a useful tool with which national statistical agencies let everyone know by how much they have missed a portion of the population; yet it can also be used to whip the agencies that are seen to have underperformed, depending on how high the undercount is.

A census is designed in such a way that everyone should be counted. However, there are always people who are missed (undercount) and people who are counted more than once (overcount). The undercount is equal to a non-response coverage and is regarded as one of the major challenges that confront many countries in census undertaking. The undercount is also often used as proxy to describe how well the count was done – in general, the lower the undercount, the better the quality of the census coverage; the higher the undercount, the poorer the coverage. Therefore, the undercount figures provide users with an assessment of the completeness of census counts that can be taken into account when using census information; especially as public concerns regarding data quality is increasing.

3.4 The Chi-square Automatic Interaction Detection (CHAID) technique

The undercount rates produced in the PES for the purposes of adjusting census data, if required, may result in skewed results when disaggregated by geography type or demographic variables. This is due to the fact that households and persons are not evenly missed over different subgroups of the population (Stats SA, 2009). The CHAID technique is used to obtain the adjustment classes by determining the combinations of predictor variables which are statistically significant in modelling the coverage (undercount) probability.

When broken down (disaggregated) by geographic or demographic variables such as province, gender, age group or population group, the overall coverage estimates could be skewed due to the fact that persons and households are not evenly missed over such subgroups of the population. Homogeneous adjustment classes, that is, classes within which coverage rates are roughly similar, are thus formed and a single adjustment factor is then calculated in each of the adjustment classes independently. The adjustment classes are obtained using the technique on PES data.

The CHAID technique is used in this investigation to determine combinations of the predictors that are statistically significant in modelling the coverage probability. CHAID works particularly well with a large sample size as it uses the chi-square to determine multiway (more than two) splits per node. CHAID also works well with categorical targets and avoids over-fitting the data. The characteristics defined by CHAID branches are then taken as the adjustment classes; thus, CHAID creates the different branches in the dendrogram (decision tree). Adjustment factors are calculated independently for each adjustment class. The predictor variables used for households are province, geographic type and household size. The predictor variables for persons are geography type, gender, age group and population group. The dependent (target) variable is defined to be the "matched population" variable (a matched person exists on a census and a PES questionnaire and is correctly counted on both), where "0" indicates non-matched, "1" indicates matched

(non-mover and out-mover) and a rate between "0" and "1" represents the probability of being matched as an in-mover.

4. CENSUS UNDERCOUNT IN SOUTH AFRICA

4.1 Causes of undercount in South Africa

The causes of an undercount can manifest in a number of places along the steps of the statistical value chain of census activities. According to Stats SA (2009), these factors include:

- Poor or ineffective planning, resulting in the lack of integration of activities especially at collection level where the omission of persons and households occurs;
- 'Hard to count' groups such as mobile populations, children, illegal migrants and even persons with disabilities;
- Limited monitoring and evaluation systems as well as quality control measures to ensure coverage;
- Lack of ownership caused by not adequately involving stakeholders in all phases of the census; ownership and buy-in improves with involvement;
- Gaps in the demarcation of the country into Enumeration Areas (EAs), resulting in some areas not being enumerated;
- Education and language barriers to filling out census forms with limited questionnaire translations and low literacy levels;
- Concerns that confidentiality of data will not be ensured due to distrust of government or a more general reluctance to participate;
- Fears over crime and high levels of distrust in some neighbourhoods, resulting in residents who fear opening their doors to strangers;
- Households were unavailable, working or travelling and difficult to contact;
- Household members mistakenly left out of the completion of the questionnaire (e.g. young babies, the elderly, or visitors);
- Dwellings missed due to location in a remote or non-residential area, or mistakenly classed as unoccupied;
- Poor questionnaire design due to poor preparation to inform the lay-out and questions to be asked, and
- Poor recruitment and training of fieldworkers.

4.2 Measurement of the undercount in South Africa 2011

Information on the undercount serves two purposes; the data is used to calculate adjustments and provide corrected estimates of the total population, and to evaluate the census coverage. The post-enumeration survey is used to establish the undercount and as a basis for weighting and correcting the raw census results. The challenge in using the PES is that, as a survey, it works well on a large scale such as on a national and provincial level; however it cannot cover local areas in detail. Thus, the challenges of 'hard-to-count' populations persist for both census and PES teams, thereby replicating bias (Christopher 2009: 107). High security walls and other security measures such as dogs in the respondents' yards discourage enumerators for both the census and the PES.

When one considers how important census data is for planning, informing interventions and information-led development, whether by private and public sectors or by non-governmental organisations, reducing the undercount is a significant priority. Furthermore, undercount affects decision-making, especially with government planning, and can thus have serious implications for the allocation of financial resources. It further affects decisions by businesses in terms of location of the businesses and offices as well as positioning products and services (Nguyen, 2007).

> In a regionally diverse, unevenly skilled, ethnically heterogeneous country like South Africa, and with a degree of political suspicion towards the new government in different areas or among different groups, the proportions of people not enumerated will vary in intricate ways (Orkin, 1998: 10).

In South Africa, as in many other countries, the rates of undercount can vary significantly for different population groups depending on factors such as gender, age, ethnicity and geographic location. In most countries, the undercount normally affects the economically disadvantaged (Citro & Cohen, 1985; Ericksen & Kadane, 1985: 104). However, in South Africa, the undercount is noted markedly among the White population, still considered economically advantaged. South Africa has different settlements in both rural and urban areas that provide different challenges as far as enumeration is concerned (Stats SA, 2009). These groups have been categorised as 'hard-to-count' and include homeless people, people living in high-walled areas (gated communities and residential complexes), informal settlements, remote rural communities and commercial farms.

Stats SA (2009) also lists the following categories as 'hard-to-count' individuals, communities or areas. The layout of industrial and commercial areas makes it difficult to identify whether people are residing on the premises. Collective living quarters such as hostels, hospitals, prisons, old age homes, etc. require a separate type of enumeration and often provide challenges to secure interviews with the correct personnel. High-walled areas are notorious for refusals and challenges to access, as are farm areas where farmers' refusal to give access to their farms is often based on security concerns and unwillingness to cooperate with the government; this often directly translates into failure to count farmworkers residing on the farm. Traditional communities can be inaccessible and require protocols to be observed. Informal settlements are usually makeshift and temporary in nature, often politically unstable and home to many diverse groups from migrants to criminals. Undocumented migrants often also refuse to participate in census activities for fear of being deported or arrested, and often do not speak local languages. Another group of people that are easy to miss are the homeless and the transient, as they are so mobile.

4.3 Undercount results for South African census years 1996-2011

4.3.1 Descriptive results

The results showing the undercount by geography type are based on calculations performed on the actual census data after the implementation of the adjustment factors sourced from the PES. The adjustment factors are derived from the undercount rates generated in the PES and are then applied according to the adjustment classes produced using the CHAID technique. The results for the 1996-census will be presented in terms

of urban and non-urban areas which do not show a more detailed picture of the different settlement types. The 2001-census geography type had four classifications, namely urban formal, urban informal, tribal or traditional area, and rural formal. It should be noted that the rural formal geography type includes commercial farms which have always proven to be the biggest source of undercount in South Africa's democratic censuses.

The 2011-census geography type was revised to include three types, namely urban, tribal or traditional, and farm. The urban geography type is a combination of both urban formal and urban informal when compared with the geography type of the 2001-census. Farms are distinctly classified in the 2011-census compared to the 2001-census where they were classified under rural formal. Table 2 shows the undercount by province and geography type for the 1996-census.

Table 2 shows that there was scant difference between urban and non-urban areas whereby the undercount rate for urban areas was 10.45% compared to 10.96% for the non-urban areas. Table 3 shows the national undercount rate by different geographic types for the 2001-census.

Table 3 shows that the rural formal geography type, which consists mostly of commercial farms, was the biggest source of undercount in the 2001-census, with 37.86%, and had the lowest percentage of the population at 7.05%. The urban informal geography type, which includes squatter camps, was the second biggest source of undercount in the 2001-census, with 24.49%, and had the second lowest percentage of the population at 7.95%. Urban formal areas had the third lowest undercount rate with a percentage of the population at 49.25. Tribal areas had the lowest undercount of 13.90%, coupled with a percentage of the population at 35.75. Table 4 shows the undercount rate by province and geography type for the 2001-census.

Table 4 shows that the rural formal areas, which consist mainly of commercial farms, are consistently the biggest source of undercount for all provinces, except Limpopo and Free State, where the urban informal areas had the biggest undercount of 56.43% and 44.30%, respectively. Urban formal areas and tribal areas had relatively lower undercount rates compared with urban informal and rural formal areas. Table 5 indicates the undercount by province and geography for the 2011-census. The Western Cape does not have any tribal or traditional areas.

Table 5 shows that farm areas had the highest national undercount of 38.14%. Urban areas had the second highest national undercount of 14.43%. Tribal or traditional areas had the lowest national undercount of 11.39%. It is also evident that farm areas retained the highest undercount in all nine provinces and that urban areas had the second highest undercount for seven provinces, except Eastern Cape and North-West. Tribal or traditional areas had the lowest undercount in seven provinces, except

Table 2: Undercount rate by province for urban and rural locations in 1996

	Derived undercount		
Province	Urban	Non-urban	Total
Eastern Cape	11.45	10.06	10.58
Free State	8.92	8.39	8.75
Gauteng	9.93	9.52	9.91
KwaZulu-Natal	12.91	12.66	12.76
Limpopo	12.92	11.07	11.29
Mpumalanga	11.74	9.44	10.23
North-West	8.00	10.24	9.47
Northern Cape	11.22	19.72	14.59
Western Cape	8.89	7.17	8.69
Total	10.45	10.96	10.69

Source: This table was calculated by R. Maluleka, Stats SA 2013

Table 3: Undercount by different geographic types in 2001

Geography type	Number of persons	Percentage of persons	Undercount rate
Urban formal	22 073 941	49.25	14.60
Urban informal	3 561 524	7.95	24.49
Tribal area	16 023 477	35.75	13.90
Rural formal	3 160 837	7.05	37.86
Total	44 819 778	100.00	16.77

Source: This table was calculated by R. Maluleka, Stats SA 2013

Table 4: Undercount rate by province and geography type in 2001

Province	Derived undercount				
	Urban formal	Urban informal	Tribal area	Rural formal	Total
Eastern Cape	12.09	24.82	12.87	30.72	14.01
Free State	8.21	44.30	7.90	42.99	16.67
Gauteng	16.52	17.80	13.62	48.04	17.46
KwaZulu-Natal	20.67	28.36	19.18	32.17	21.42
Limpopo	9.98	56.43	10.80	52.48	14.23
Mpumalanga	7.69	8.79	10.92	45.96	14.83
North-West	15.73	16.02	12.69	29.78	15.88
Northern Cape	9.76	8.13	12.19	30.68	12.99
Western Cape	12.80	25.68	-	26.72	15.15
Total	14.60	24.49	13.90	37.86	16.77

Source: This table was calculated by R. Maluleka, Stats SA 2013

Table 5: Undercount by province and geography type in the 2011-census

Province	Derived undercount: Census 2011		
	Urban area	Tribal or traditional area	Farm area
South Africa	14.43	11.39	38.14
Western Cape	14.87	-	45.93
Eastern Cape	10.38	12.48	43.91
Northern Cape	16.68	4.82	23.64
Free State	9.33	4.95	23.71
KwaZulu-Natal	17.88	14.15	30.62
North-West	11.98	12.04	54.12
Gauteng	14.28	12.95	34.92
Mpumalanga	16.98	11.10	34.60
Limpopo	17.85	7.70	34.40

Source: This table was calculated by R. Maluleka, Stats SA 2013

Eastern Cape and North-West. There are no tribal or traditional areas in the Western Cape.

The three censuses conducted in post-apartheid South Africa all seem to point to consistent challenges which are directly linked to the different geography types, particularly farm areas and informal settlements. Each census is meant to compile lessons learnt and future improvements that should result in targeted interventions necessary to reduce undercount in the worst performing geography types.

The results for the 2001 and the 2011 censuses clearly show that there was virtually no improvement in the enumeration of farm areas. In both censuses, farm areas have the highest undercount rate at national and provincial levels. The challenges faced by census staff in enumerating farm areas are well documented. The national undercount rate of farm areas in the 2011-census was 38.14% compared to 37.86% recorded for rural formal areas in the 2001-census. This is a clear indication that the interventions for the 2011-census did not yield the required reduction of undercount in farm areas.

The relatively unchanged and unacceptably high undercount rate from 2001 to 2011 implies that there are recurring challenges and obstacles in the enumeration of farm areas. These problems may be organisational and internal to Stats SA in terms of having a clear and more robust strategy for enumerating farm areas in addressing the issues of access to farms through engagement of farmers' unions, suitable vehicles for rough terrain, dealing with seasonal workers, and boundary verification for the abnormally large commercial farms.

Urban areas also consistently produce the second highest undercount rate which is mainly located in informal settlements. The national undercount rate in the 2011-census was 14.43% compared with 14.60% for urban formal areas in 2001. It would be interesting to note the national undercount rate for the equivalent of urban informal in 2011; the undercount rate for urban informal stood at 24.49% in 2001 and is likely to be similar in 2011. The enumeration of informal settlements is not necessarily hampered by bureaucratic challenges similar to farms, but owes most of its challenges to the listing exercise. The lack of formal streets and numbering systems in informal settlements contribute to these challenges. Another challenge is the existence of backyard shacks, particularly those surrounding RDP houses. This complicates the identification of dwelling units and households. In some areas, the presence of foreigners, who view the census suspiciously, also leads to an increase in the undercount rate. To some extent, the choice of a *de facto* (presence at specific place and time) census versus a *de jure* (usual place of residence) census might have the biggest impact in urban informal areas, as residents find the concept of the census reference night difficult, due to local patterns of short-term migration.

Tribal or traditional areas have consistently produced the lowest undercount rates. The national undercount rate was 11.39% which is a slight improvement on the 2001 rate of 13.90%. The absence of formal streets and numbering in most traditional areas does not seem to have a significant impact on the undercount rate, as these areas have consistently fared better than farm areas and informal settlements.

4.3.2 Chi-square analysis

The results from the Chi-square analysis of the role of geography type in the formation of adjustment classes (homogeneous groups) for the 2011-census are important. Eight of the nine provinces showed similar outcomes; different results were recorded only in Gauteng. The dendograms (decision trees) below will demonstrate the outcomes for the Free State (as is for the other seven provinces) and for Gauteng. Figure 1 illustrates the decision tree for the undercount of all households, whereas Figures 2 and 3 show the decision tree for the undercount of persons in the Free State and Gauteng, respectively.

Figure 1 shows the decision tree for the undercount of all households. The Household Size Group (HHSIZE_GRP) was the most significant predictor variable in explaining the variation in the undercount. Province was the second most significant predictor variable while settlement type (urban, traditional and farm) was the least significant predictor variable. The Settlement type (EA_GTYPE_C) was less significant for households undercount than for persons undercount, showing that

- 0 = non-matched (counted in the census or PES, but not in both).
- 1 = matched (counted in both the census and the PES).
- The remainder of the codes indicated a match status other than 0 and 1, e.g. census erroneous inclusion.

Figure 2 shows the decision tree for the undercount of persons in the Free State. The decision tree shows that settlement type (EA_GTYPE_C) was the most significant variable in explaining the variation in the persons undercount.

The second most significant predictor variable was Population group (Popgroup_code), showing that

- 0 = non-matched (counted in the census or PES, but not in both).
- 1 = matched (counted in both the census and the PES).
- 0.6979 = a rate between "0" and "1" represented the probability of being matched as an in-mover.

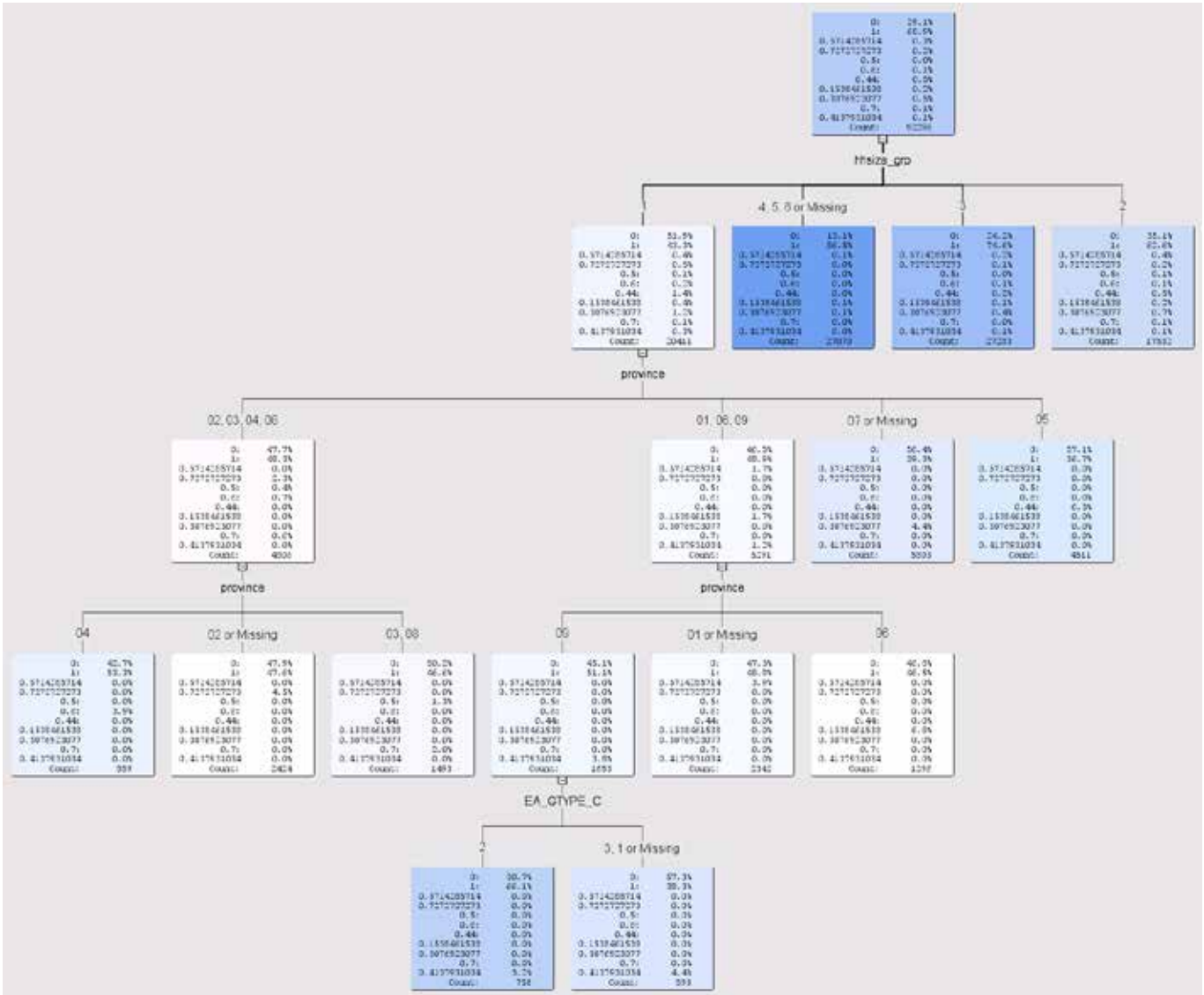

Figure 1: CHAID output for the undercount of all households: Census 2011

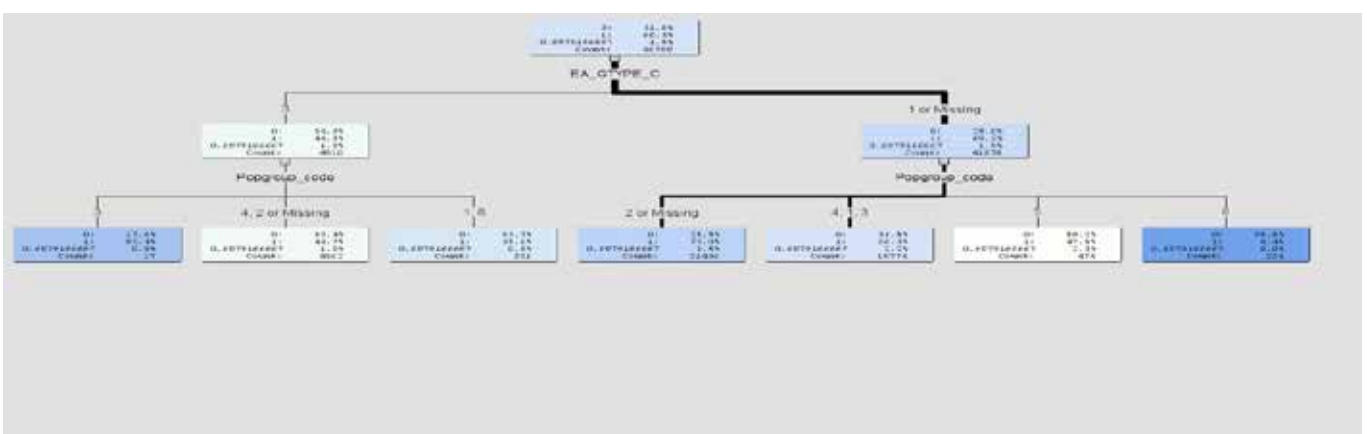

Figure 2: CHAID output for persons undercount for the the Free State: Census 2011

Figure 3 shows the decision tree for the undercount of persons in Gauteng. Unlike the other eight provinces, the most significant predictor variable for Gauteng was population group. Population group (Popgroup_code) was the most significant variable in explaining the variation in the persons undercount. It was followed by the EA_GTYPE_C, showing that

- 0 = non-matched (counted in the census or PES, but not in both).
- 1 = matched (counted in both the census and the PES).
- 0.6216 = a rate between "0" and "1" represented the probability of being matched as an in-mover.

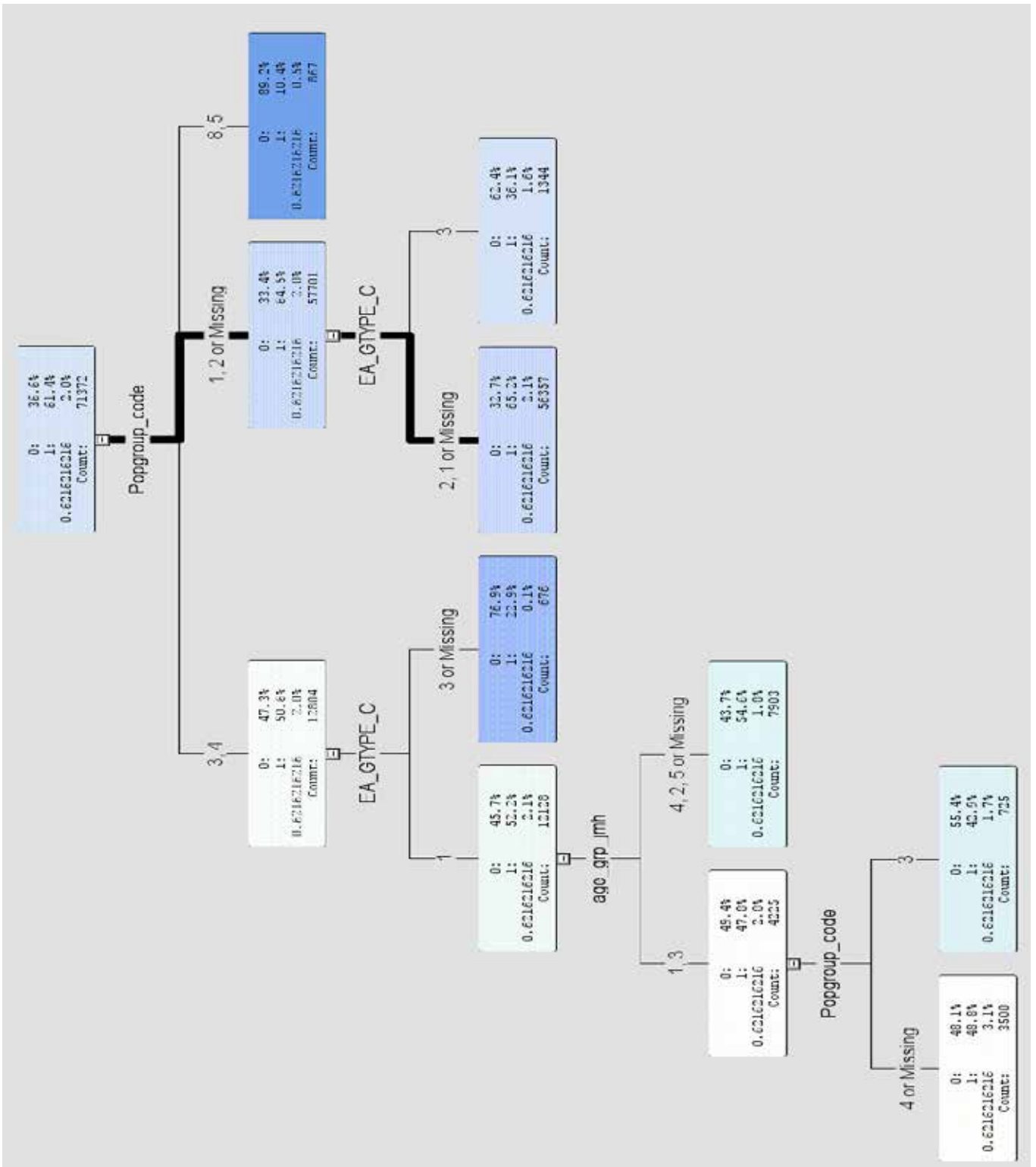

Figure 3: CHAID output for persons undercount for Gauteng: Census 2011

In conclusion, the results indicate that, with the undercount of all households, the variable household size was the most significant predictor variable in explaining the variation in the undercount for all the provinces. This differs from the undercount of all persons (except Gauteng) in that the settlement type was the most significant variable in explaining the variation in the persons undercount. In Gauteng, the most significant predictor variable was population group.

5. CONCLUSION

The results presented in this article clearly show that different geography types present contrasting undercount rates with the farm areas (rural formal areas in 2001), with informal settlements being consistently the worst affected by undercount, whereas urban formal and tribal or traditional areas always fare relatively better. This shows that geography type historically played a part in the origins of the undercount in South Africa. The CHAID output also shows that geography type plays a significant role in the origins of the undercount, particularly at person level.

After the third democratic census, it would be expected that decisive interventions in the enumeration of farm areas and squatter camps would have been prioritised; however, the results indicate that these interventions were relatively inadequate and ineffective in achieving set objectives such as a lower single digit undercount in 2011.

The undercount rate can be attributed to a weakness in the approach by Stats SA for dealing with known causes of undercount, particularly in farm areas and urban informal areas, as the three democratic censuses show that there are recurring challenges in the enumeration of farm areas and informal settlements. Another matter concerning the undercount is the willingness of the population to fully participate in the census as a non-political exercise. The aforementioned issues illustrate that more detailed studies are required by Stats SA in order to get a better understanding of the origins of the undercount and to improve all planning and operational arrangements with a specific focus to reduce the undercount in farm areas and informal settlements.

This study underlines that a clear understanding of census undercount can contribute to a higher level of confidence in census data which is crucial for effective planning, programming, monitoring and development both by public and private organisations. Future research could also confirm, through empirical analysis, more factors contributing to census undercount in South Africa. Addressing these issues can lead to an improved level of confidence in the census data which, in itself, can lead to an improved implementation of the strategies referred to in the NDP vision of 2030.

REFERENCES LIST

AHMED, B., GUPTA, P.D., ROBINSON, J.G. & WOODROW, K.A. 1993. Estimation of population coverage in the 1990 United States Census based on demographic analysis. *Journal of the American Statistical Association*, 88(423), pp.1061-1071.

AUSTRALIAN BUREAU OF STATISTICS. 2011. Estimates of net undercount. [online]. Available from: <http://www.abs.gov.au/ausstats/abs@.nsf/Products/2940.0~2011~Main+Features~Estimates+of+net+undercount?OpenDocument> [Accessed: 17 September 2013].

BELL, W.R.1993. Using information from demographic analysis in post-enumeration survey estimation. *Journal of the American Statistical Association*, 88(423), pp. 106-118.

BROWN, L.D., EATON, M.L.,FREEDMAN, D.A., KLEIN, S. P., OLSHEN, R.A., WACHER, K.W., WELLS, M.T. & YLVISAKER, D. 1999. Statistical controversies in Census 2000. *Jurimetrics*, 39, pp. 347-375.

BYCROFT, C. 2006. Challenges in estimating populations. *New Zealand Population Review*, 32(2), pp. 21-47.

CARR-HILL, R. 2012. Missing millions and measuring progress towards the millenium development goals with a focus on Central Asian states. *Central Asian Journal of Global Health*, [online], 1(1). Available from: <http://cajgh.pitt.edu/ojs/index.php/cajgh/article/view/24/30> [Accessed: 11 March 2013].

CARTER, M. 2009. Explaining the Census: Investigating reasons for non-response to the ABS Census of population and housing. Swinburne University of Technology: Institute for Social Research, pp. 1-33. Available from: <http://www.sisr.net/documents/Census.pdf.> [Accessed: 12 September 2012].

CITRO, C.E. & COHEN, M.L. 1985. *The bicentennial census, new directions for methodology in 1990*. Washington, DC: National Academy Press.

CHOLDIN, H. 1994. *Looking for the last percent: The controversy over census undercount*. New Brunswick, NJ: Rutgers University Press.

CHRISTOPHER, A.J. 2009. Delineating the nation: South African censuses 1865-2007. *Political Geography,* 28, pp. 101-109.

ELLIOT, M.R. & LITTLE, R.J.A. 2005. A Bayesian approach to 2000 Census evaluation using ACE survey data and demographic analysis. *Journal of the American Statistical Association*, 100(470), pp. 380-388.

ERICKSEN, E.P. & KADANE, J.B. 1985. Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association*, 80(389), pp. 98-109.

FAY, R.E., PASSEL, J.S. & ROBINSON, J.G. 1988. The coverage of population in the 1980 census. Evaluation and research report PCH80-E4. Washington, DC: U.S. Department of Commerce.

FRAZER, J.G. 1918. *Folklore in the Old Testament: Studies in comparative religion, legend, and law*. London: Macmillan.

FRIED, J. 1993. Settlement types and community in Northern Canada. *Arctic Institute of North America*, 16(2), pp. 93-100.

HOGAN, H. 1993. The 1990 post-enumeration survey: Operations and results. *Journal of the American Statistical Association*, 88(423), pp. 1047-1060.

ISAKI, C.T. & SCHULTZ, L.K. 1986. Dual system estimation using demographic analysis data. *Journal of Official Statistics*, 2(2), pp. 169-179.

KOCH, J. & CEBULA, R. 2004. The final 2000 census state response rates: myths and realities. *Social Science Journal,* 42(2004), pp. 575-585.

MARTIN, D. 2010. Understanding the social geography of census undercount. *Environment and Planning A*, 42(11), pp. 2753-2770.

MASITENG, K.K. & KEKOVOLE, J. 2006. Planning of Census 2011 approach and implementation strategies. [online]. Available from: <http://unstats.un.org/unsd/demographic/sources/census/2010_phc/South_Africa/South_Africa.pdf.> [Accessed: 12 January 2013].

NGUYEN, P. 2007. The census bureau and its accountability. *The American Review of Public Administration*, 37(2), pp. 226-243.

NPC (NATIONAL PLANNING COMMISSION). 2011. *National Development Plan 2030: Our future: Make it work*. Pretoria: National Planning Commission, Department The Presidency.

ORKIN, M. 1998. *Calculating the undercount in Census '96*. Pretoria. Statistics South Africa.

RAZI, B.J. 1999. Census politics revisited: What to do when the government can't count? *American University Law Review*, 48(5), pp. 1101-1138.

SAGNA (SOUTH AFRICAN GOVERNMENT NEWS AGENCY). 2012. Census undercount of 14.6% a concern. [online]. Available from: <http://sanews.gcis.gov.za/south-africa/census-undercount-146-concern> [Accessed: 17 September 2013].

SIMON, P. 2012. Collecting ethnic statistics in Europe: A review. *Ethnic and Racial Studies*, 35(8), pp. 1366-1391.

STATISTICS NEW ZEALAND. 2007. *A report on the 2006 post-enumeration survey*. Wellington: Statistics New Zealand.

STATS SA (STATISTICS SOUTH AFRICA). 2009. Census undercount and strategies: Version 6. Pretoria: Statistics South Africa.

UK STATISTICS COMMISSION. 2004. *Census and population estimates and the 2001 census in Westminster: Final Report*. London: Statistics Commission.

UN (UNITED NATIONS). 2001. Handbook on census management for population and housing censuses. Series F No. 83/Rev1. New York: Department of Economic and Social affairs Statistics Division.

US CENSUS BUREAU. 2000. US Census Bureau News: Statement by William G. Barron Jr. on the current status of results of census 2000 accuracy and coverage evaluation survey US Census Bureau, Washington [online] Available from < http://www.census.gov/newsroom/releases/archives/2010_census/cb12-95.html> [Accessed: 16 January 2013].

WRIGLEY, E.A. & SCHOFIELD R.S. 1989. *The population history of England, 1541-1871: A reconstruction*. Cambridge United Kingdom: Cambridge University Press.