SAJDem 9(1):67–90

Comparing Census 1996 with Census 2001: An operational perspective

Marius Cronje and Debbie Budlender[1]

Abstract
Statistics South Africa conducted the country's first post-apartheid population census in October 1996. Exactly five years later, the agency conducted a second population census. This paper compares the two censuses from a field operations perspective. The aim is to help users of the data judge whether particular observed differences between the two censuses are due to real changes in the population or to changes in methodology or quality of the enumeration. The paper describes the overall approach adopted in the two censuses, the questionnaires and topics covered, demarcation and listing, field operation structures and training, the pilots, enumeration procedures, processing, and the post enumeration survey. The paper points to both weaknesses and strengths in the changes effected between the two censuses, and in how they were implemented.

Key words
South Africa, census, field operations, demarcation, post enumeration survey, questionnarie design, enumeration, listing, fieldworker training

Introduction

A population census is a major undertaking in any country. It requires enormous financial, human, and other resources. The expenditure of the resources is considered worthwhile because the data provided are used to inform policy makers about the life circumstances and needs of the population. The census also provides a base for statistics, and a statistical frame for sampling surveys and studies between censuses.

1 Marius Cronje, Statistics South Africa, Private Bag X44, Pretoria, 0001. Email: *mariusc@statssa.gov.za*
Debbie Budlender is Senior Researcher at the Centre for Actuarial Research at the University of Cape Town.

Statistics South Africa is the official agency responsible for conducting national population censuses in South Africa. The Statistics Act of 1999 requires that Statistics South Africa does this every five years, despite the fact that ten-year censuses are the norm. Other countries that follow a 5-year census cycle include Australia and Canada.

The first post-apartheid census was conducted in October 1996. In October 2001, Statistics South Africa conducted the second population census since the country's first democratic elections of 1994. One of the many challenges for users of the census data is to draw comparisons between the two censuses, and to be able to say which of the observed differences are due to real changes in the population and which are due to changes in methodology or quality of the enumeration.

This paper compares the two censuses from a field operations perspective. It describes the procedures used and highlights changes between the two censuses. It discusses, in particular, the overall approach, questionnaires and topics covered, demarcation and listing, field operation structures and training, the pilots, enumeration procedures, processing, and the post enumeration survey (PES).

Overall approach

In both censuses, the reference point was the night of 9–10 October. Both censuses were conducted on a *de facto* basis i.e. they enumerated each person at the place where they were at midnight on 10 October, regardless of their usual place of residence (Central Statistical Service 1996b; Statistics South Africa 2001b).

In the apartheid years, different approaches were used for enumeration in different areas. In particular, some 'black' areas were "enumerated" by means of estimates from aerial photographs as it was considered too dangerous for enumerators to go door-to-door. The 1996 census was the first attempt to standardise methodology for all areas, and this practice was repeated in 2001.

In both censuses, enumerators visited every household and either interviewed a household representative, or left the questionnaire for the household to complete. In both censuses, face-to-face interviewing was regarded as the default, and the household was meant to complete the questionnaire themselves only if they insisted on doing so (Statistics South Africa 2001b:83). This approach was adopted in light of the high levels of illiteracy in the country. A further advantage is that consistency in the asking

(and interpretation) of questions should have been enhanced, since enumerators had been trained to ask (and understand) questions in particular ways. The fact that respondents are allowed to fill in the questionnaires themselves holds the danger that some of the questions might be misinterpreted.

Enumerators were required to indicate on the questionnaire whether they or the respondent completed the questionnaire. An initial assessment of the 2001 census data suggests that approximately four per cent of questionnaires were completed by respondents. In 1996, too, a small minority of questionnaires were completed by respondents.

Questionnaires and topics covered

Translations

For both censuses the questionnaires were translated into all 11 official languages. However, the way in which translations were made available differed between the two censuses. For Census 1996, all 11 versions of the questionnaire were available in the field. For census 2001, every enumerator had a translation booklet that could be used by the respondent to read the questions in their preferred language. The responses would, however, usually be captured on an English version of the questionnaire although the appropriate actual questionnaire was also available on request. Unfortunately, enumerators were not required to record in which language the interview was conducted and there is no way of checking whether this might have affected responses.

These changes were introduced to ease the logistical difficulties involved in distributing all the different questionnaires. The changes could, however, have caused some decrease in quality. In particular, reading questions from the translation and filling them in on the English version could have resulted in errors.

In fact, Statistics South Africa discovered some errors in the translations subsequent to the 2001 census. For example, in one question some of the translations had one of the options missing. Further, some of the more complicated options – for example, in respect of area of post-school education – were not translated into some languages due to difficulties in finding suitable words for the concepts.

Number of questionnaires and questions

Four different questionnaires were used in 1996, while three were used in 2001. In 1996, the primary questionnaire was a household questionnaire,

which was used for private households and for hostels which provided family accommodation. It contained 20 questions about each person and 8 for each household. In 2001, the equivalent questionnaire was used for private households, as well as for worker hostels, student hostels, homes for the aged, and residential hotels (Statistics South Africa 2001e; Statistics South Africa 2001b). It contained 22 questions about each person, and 9 questions about the household as a whole. This questionnaire was thus similar for both censuses in whom it reached and the number of questions, except that in 2001 its use was extended to additional types of living quarters.

The second questionnaire used in 1996 was a summary book for hostels which listed all persons/households in the hostel as well as containing 9 questions about the hostel as a whole. A third questionnaire book was used in 1996 to record the answers to 19 questions on individuals within the hostels. These 19 questions were identical to those in the household questionnaire with the exception of the relationship to household head question, which was omitted for obvious reasons. The fourth questionnaire in 1996 was used for "special enumeration". It was used to record responses to six basic questions about individuals within institutions such as hotels, prisons, hospitals etc. as well as for homeless persons. The questionnaire also included nine questions about the institution as a whole.

In 2001, the second and third questionnaires were used to cover all institutions other than worker hostels, student hostels, homes for the aged and residential hotels. Questionnaire B contained 21 questions about individuals, while questionnaire C contained 8 questions about the institution as a whole. Again, the question on relationship to household head was omitted on the personal questionnaire. Questionnaire B was also used to enumerate homeless people (Statistics South Africa 2001b:78).

The changes in the questionnaires were designed to obtain more easily comparable information about all individuals, whether resident in private homes, hostels or other institutions. In particular, the changes provided for the collection of more comprehensive information on individuals in institutions.

Topics covered

Both censuses covered a range of demographic, social and economic topics. Some questions applied to each individual in the household, for example, respondents were asked to indicate the age, sex, population group, home language, education level, occupation and income of each person present in

the household on census night. Other questions dealt with the circumstances of the household as a whole, for example, whether or not the household had access to electricity and piped water.

For Census 2001 there was a co-ordinated effort to try and harmonise census questions across the Southern African Development Community (SADC) region so as to have comparable data. A core set of questions was agreed to by member states in the SADC region, and Statistics South Africa included all these core questions with the exception of questions on the material used for the construction of roofs, walls, and floors. It then supplemented the core questions with others relating to identification of the individual's spouse, population group, language, religion, movement since Census 1996, field of highest post-school qualification, work-seeking activities, hours worked in previous seven days, work location, mode of travel to school or work, income, sharing of a room with another household, main source of water for domestic use, household items in working order, and telephone usage (Adegboyega 2001:6–10).

Overall, the questionnaires for 2001 were kept as similar as possible to those from 1996 to facilitate comparison of changes in the South African population over time. The differences between the two years in terms of questions added or omitted can be summarised as follows.

Questions absent in 1996 but added in 2001

- Which member of the household was the spouse
- Province of birth of all individuals
- Whether an individual took active steps to find employment in the four weeks before the census
- Hours worked in the seven days prior to the census
- Date, sex and vital status of last child born
- Travel to school or work
- Type of living quarters
- Whether there is more than one dwelling on the site
- Means of obtaining piped water
- Household goods (radio, refrigerator, television, telephone in the dwelling, computer, cell-phone) in working condition
- Deaths in the household in past 12 months

Questions asked in 1996, but omitted in 2001

- Whether the respondent was a migrant worker
- Previous activities of unemployed members of the household
- Whether an individual worked full- or part-time
- Main duties or activities performed at work
- Additional income
- Date of first child born
- Numbers of births in the twelve months before the census

The number of questions implied in the list of changes above may seem to contradict the number of questions for the questionnaire of each year cited above. This is explained by the fact that some questions have multiple parts. The number of questions in the earlier part of the paper counts each part separately. The list above refers to topics, and some of these topics are covered by multiple-part questions.

Some of the changes in 2001 provide for coverage of new topics, such as deaths in the household over the past 12 months. The inclusion of this question in 2001 might prove a useful indicator of the impact of HIV/AIDS on the country. The additional questions in respect of access to water reflect the intense governmental interest in this area of service delivery. The change will, however, make comparisons of access to water between the two years difficult.

The additional question on attempts to find work in the past four weeks will provide for a clearer distinction between the official measure of unemployment and the expanded definition, which does not require a person to have looked for work in the past month to be classified as unemployed. The ability to derive the official rate of unemployment from Census 2001 was intended to allow for more accurate comparison with the results of subsequent rounds of the six-monthly Labour Force Surveys. However, analysis of the data from Census 2001 revealed that the way the initial question on work during the past seven days was framed resulted in many individuals working in agriculture or the informal sector not being reported as having worked. This, in turn, affected the unemployment rates as the number of employed people forms part of the denominator. The rates from Census 2001 and the Labour Force Surveys are thus not easily comparable.

The dropping of the question about duties or activities at work will result in a deterioration in the quality of occupational information in Census 2001. Occupation is one of the "write-in" questions in both years, and therefore has to be post-coded. The occupational title alone often does not provide

sufficient information for accurate coding, and internationally it is accepted that a second question, asking about tasks and duties, is useful to provide the further information needed. This information will not be available to those responsible for coding Census 2001. The question about tasks and duties was dropped in an attempt to keep the questionnaire (relatively) short. This decision may need to be reconsidered for future censuses if Statistics South Africa wants to provide quality information about occupation.

In addition to the inclusion and exclusion of entire questions, further changes were made to the way questions on particular topics were phrased and/or the options offered in terms of responses. For example, the list of options for marital status was expanded in 2001 to distinguish polygynous marriages from other customary marriages. Divorced and separated individuals were also coded separately in 2001. In time, analysis will reveal how well these distinctions were understood and reported.

Some questions were expanded in 2001, while others became less comprehensive. Language is one of the areas of contraction, in that in 2001 the questionnaire enquired only about the language spoken most often in the household, while the 1996 questionnaire also enquired about further languages spoken at home. The citizenship question in 2001 asks only whether the person is a South African citizen or not, while in 1996 there was provision for reporting dual citizenship. It is, however, debatable whether this question was answered either fully or accurately in 1996 as many respondents might have been uncertain as to the legality or otherwise of dual citizenship.

The number of questions asked on women's fertility, on the other hand, was increased significantly in 2001, while the phrasing of the questions themselves was also changed. However, evidence from the pilot census suggests that the increase in the number of questions, while increasing the information potentially available, also increases the potential for contradictions between the responses to different items offered in respect of a particular woman. These contradictions pose challenges for the cleaning process which attempts to make data consistent.

Questions in respect of post-school education were also changed between 1996 and 2001. In Census 1996, respondents were asked to name the highest qualification and the question was a "write-in" one which had to be post-coded. Those responsible for the post-coding experienced considerable difficulties because of the ways in which respondents reported qualifications. The 2001 questionnaire asked respondents to mark off which of 22 learning

areas is the relevant one for best describes the person's highest post-school qualification. This approach has the advantage of avoiding post-coding. The disadvantage is that while this categorisation is used by all universities and technikons in their annual reporting to the Department of Education, it is not widely known by the public. Further, the names of the learning areas are not always intuitively easy to understand. One indication of the difficulties in this respect is that some of the translators of the questionnaires were not able to find appropriate translations in their language and left the names in English. The change in the way this question was asked might have created difficulties for respondents and enumerators in completing the question in Census 2001.

In both years, the questions in respect of personal income were framed in terms of income categories rather than asking for the exact income of each household member. This approach was adopted both because people do not always know the exact income, and because it is felt that on this sensitive question, respondents may be more willing to give an approximate than an exact response. The cut-off points for the categories differed between the two years. This was more or less inevitable, given that inflation had not been negligible in the intervening five years. Unfortunately, however, the new cut-offs are not simply an inflation adjustment of the previous ones. Comparison of the results of the two years in respect of income will thus be difficult.

Demarcation and listing

Different approaches to the demarcation of enumeration areas (EAs) and listing of dwellings in each EA were adopted in the 1996 and 2001 censuses. In 1996, demarcation and listing were done simultaneously, whereas in 2001 they were done separately. GIS resources were used extensively in 2001, but only to a limited degree in1996. More than 200 000 maps were generated for census operations in 2001 (North 2002).

Enumerator area types

In 1996 the country was divided into 86 000 EAs, which were grouped into 15 EA types (see CSS 1998a,b for full details). The EA types were defined, firstly, by the geographical location, and, secondly, by the type of dwelling that predominated within the EA, for example formal dwellings, informal dwellings, hostels, or institutions. In terms of geographical location, the grouping distinguished between:

- EAs in an urban municipal area, i.e. an area within a proclaimed local authority;
- EAs in an area adjacent to an urban municipal area; and
- EAs in a non-urban (rural) area not adjacent to an urban area.

These three groupings resulted in what were commonly referred to as "urban", "semi-urban", and "non-urban". The "non-urban" was sometimes also termed "rural", although this was not strictly accurate. The semi-urban, which accounts for only around 3 per cent of the population, was sometimes grouped with urban and sometimes with non-urban, but usually with non-urban.

The following 10 types were used to identify the EAs for Census 2001 (Statistics South Africa 2001c:6):

- Vacant
- Tribal settlement
- Farms
- Small holdings
- Urban settlement
- Informal settlement
- State park
- Industrial area
- Institution
- Hostel

In addition, three sub-types (no institution, institution only, and mixed) were created to accommodate institutions that were too small to be an EA on their own (Statistics South Africa 2001f:46). These sub-types were to be demarcated as part of the surrounding EA, with the sub-type reflecting the difference in their situation.

Number and distribution of enumerator areas

The number of EAs decreased between 1996 and 2001, from 86 2000 to 80 782. This was due to the fact that some EAs that were demarcated in 1996 were combined to form EAs in 2001, while other EAs were divided into more than one EA. Combination occurred in cases where vacant EAs in 1996 were combined with non-vacant EAs and where institutions that were considered as EAs on their own in 1996 were combined to form part of other EAs. The decrease was not equally distributed across the provinces. The number of EAs in the Eastern Cape increased from 16 100 to 18 371, while the number in

Northern Cape increased less dramatically from 1 500 to 1 661. All other provinces saw a decrease in the number of EAs. The decrease was most marked in Gauteng – from 17 100 to 13 367.

Recruitment

In 1996, most fieldworkers were recruited from the ranks of the unemployed. This approach was chosen in light of the relatively high unemployment rate in the country which meant, firstly, that there were large numbers of unemployed people with the required level of education (completion of grade 12) and, secondly, that there was a need for spreading income-earning opportunities. The approach differed from that adopted in South Africa in the apartheid years and in many other countries where people who already have jobs – teachers, in particular – are given census enumeration as a second job.

The approach adopted in 2001 was much the same. The focus was on recruiting people from the areas where they were to work. This task proved to be a big challenge, especially in the more affluent areas where residents did not find the job of enumerator attractive. Regional offices tried to ensure that they received applications from all the areas in their jurisdiction. If there was a specific area that was not covered, they made a special effort to get applications from these areas by talking to community organisations and leaders (Statistics South Africa 2001d:6–7).

In both censuses, Statistics South Africa endeavoured to ensure that people were working in areas with which they were familiar. This meant, firstly, that the demographic characteristics of the fieldworkers were simular to those of people living in the area. Secondly, it was argued that if enumerators were familiar with the area that they worked in, the work would be easier. Out of concern with confidentiality, people were not appointed to act as enumerators in the exact EA where they stayed but rather in an area nearby.

Training

In 1996, training of fieldworkers was organised according to a cascade approach. Fieldworkers (enumerators, chief enumerators and supervisors) were trained for three days (CSS 1996d:5). The training was mainly theoretical and conducted in a classroom format. The training for Census 2001 was significantly different from that of Census 1996. The 2001 census started with a 9-day intensive training course for regional staff. Fieldwork co-ordinators and regional trainers were trained as one group. Because of the

number of trainees, training had to be conducted in two rounds for fieldwork co-ordinators and regional trainers. Classes had to be small enough, no more than 30 people, to allow the use of live transmissions from Pretoria. The training venues were spread throughout the provinces. The transmissions used a technique called narrow casting, whereby Statistics South Africa in Pretoria was given a unique frequency to use and all the training venues tuned in to that frequency.

Subject experts were used at the centre in Pretoria. They engaged in an interactive process in which trainees would watch the transmission and then compile questions that were phoned, faxed, or e-mailed to the transmission centre in Pretoria. The experts' answers to these questions would then be transmitted to all trainees. The trainees were also assisted in the training rooms by national trainers. This training lasted for nine days. The fact that experts from head office were used to do the training meant that the correct message went out to all the trainees at the same time from the same person. The training sessions were taped and used in the training of other fieldworkers.

The next round of training was aimed at supervisors and lasted for six days. The training was a combination of in-class training combined with training videos and practical exercises. Regional trainers assisted by fieldwork co-ordinators conducted the training. The national trainers and head office staff acted as monitors in this process.

The last round of training was aimed at enumerators. It was conducted in three sessions of six days each so as to be able to accommodate the large number of trainees while still being able to use videos and keep classes small. The training comprised both theoretical and practical aspects. The supervisors assisted by the regional trainers and fieldwork co-ordinators conducted the training.

The major accomplishment for the training of fieldworkers for Census 2001 relative to Census 1996 was that it was uniform. Another improvement was that map-reading skills were included in this training. This aspect was lacking in the previous census.

Manuals

The manuals used for fieldwork training were much more detailed in Census 2001 than in Census 1996. Further, the 2001 manual included many examples as well as a range of practical exercises. In particular, supervisors and enumerators were given detailed training on map reading and interpretation

(Statistics South Africa 2001f; Statistics South Africa 2001b:17–21), which was lacking in Census 1996.

Pilot

Both censuses were preceded by pilots. These were intended to test all the systems required to run a full census, and address issues of operational difficulty. After both pilot censuses, a major debriefing was conducted. These sessions formed the basis of any changes that had to be made.

The pilots were intended primarily to test operational procedures in the field. Unfortunately, in 2001 the pilot was conducted later than originally planned because approval of the necessary budget was obtained later than expected. The relatively short period between the pilot and the census proper meant that there was not sufficient time to input and analyse the responses and make any changes to the questionnaire or operations which these processes might have suggested were advisable. The pilot data were, however, captured and used in the development of a prototype editing programme.

For Census 2001, changes post-pilot included some involving the questionnaire. In some cases the translations (in the translation booklets) were not changed accordingly. Other post-pilot changes involved logistical and operational issues such as distribution of the material to the field, the content of the enumeration kit, and some administrative procedures, for example the forms to be completed. A risk for both censuses was that some of the changes effected after the pilot censuses could not be tested properly before being implemented in the main census.

Enumeration procedures

There were some significant changes in enumeration procedures between 1996 and 2001. On the cosmetic side, the first change was brought about by the need to ensure access to households so that interviews could be conducted or questionnaires deposited. To facilitate access, steps were taken to ensure that enumerators, supervisors, and regional office staff could be clearly identified as visiting on official business. The 2001 census logo changed slightly in format and colour but was intended to build on the brand recognition that had developed around Census 1996.There were, however, still some significant difficulties in obtaining access in more affluent areas where dwellings are protected by high security walls. Problems were also experienced in obtaining access to farm workers living on commercial farms.

On the administrative side, Census 2001 saw a reduction in the administrative duties of all levels of staff involved in the field operation. This was intended to free up enumerators, supervisors and office staff so that they could concentrate on the completion and quality checking of the questionnaires. The number of administrative forms for enumerators and supervisors was reduced and the administrative duties fell primarily on the field work co-ordinators (CSS 1996d; Statistics South Africa 2001d). However, even with the reduction of the number of forms, there were still complaints about the burden on the fieldworkers.

Despite the attempted improvements in operational structures and training and reduced administrative loads, problems were inevitably encountered during implementation. These problems resulted in slower enumeration than originally planned, so that the enumeration period had to be extended several times. The extension of the enumeration period had a range of implications. In addition to delaying later procedures and increasing the costs of enumeration, it could have affected the quality of the information collected. In particular, the delay meant that that there was a longer gap between the date for which the census attempted to collect information (night of 9/10 October) and the date on which this information was collected. The bigger the gap between 10 October and the date of enumeration, the greater the likelihood that the people present on the night of 9 October were no longer resident in the household, and the greater the likelihood that respondents would have forgotten the exact situation that pertained on the night of 9/10 October.

Processing

Census 2001 saw a significant increase in the use of sophisticated technology in different stages of the census. This includes the increased use of GIS described above, the use of video in training, and barcodes for questionnaires. There was also an increased use of sophisticated technology at the processing stage.

The data-processing phase involves the conversion of information collected on the questionnaires during the enumeration phase into electronic format. Coding of open-ended questions such as occupation, data capture, checking and editing are all part of this phase.

In Census 1996 the processing took place at nine sites, one in each province. It occurred throughout 1997 and part of 1998. Approximately 5 000 temporary staff members worked in three shifts to complete the task. For Census 2001 one site was used, with approximately 800 people working

in three shifts. The dramatic decrease in the number of staff employed was a result of the shift from a largely manual operation to an electronic document imaging solution. Processing started in November 2001, and was completed in May 2003.

In Census 1996, the first step in processing was for coders to manually fill in the appropriate code for the open-ended questions onto the questionnaire after consulting a manual and to then capture these data on computer. The data then went through an editing and data verification process that attempted to remove inconsistencies. This was done with the help of custom-made software developed by Statistics South Africa. These steps all took place in the provinces. The next step, data integration, involved bringing the data from the provinces to Statistics South Africa's head office and running additional data quality checks. This step also involved additional editing programmes, the combination of the different data sets and the development of derived variables. This was done by experts at head office.

For Census 2001, the paper documents which constitute the questionnaires were converted into electronic images and data. This was done through scanning the questionnaires and extracting data with a combination of the following techniques:

- Optical mark recognition (OMR) for tick boxes;
- Optical character recognition (OCR) for machine-printed characters such as bar codes; and
- Intelligent character recognition (ICR) for hand-written alpha-numeric characters.

The use of scanning required new methods of handling the documents as, in scanning, the pages of each questionnaire need to be separated. To avoid "orphaned" pages, each page was allocated the bar code unique to a particular questionnaire so that it could be traced back to the correct questionnaire (Procon 2001).

Open-ended questions, such as occupation, were captured through computer-assisted coding. For occupation, for example, the coder allocated a code by referring to a manual the first time a particular occupation was recorded, but each further occurrence of an identical entry was allocated the same code automatically. This approach was expected to be much faster than the previous manual coding. However, the process still took far longer than expected. To address this problem, a further level of automation was developed in which the computer automatically allocated codes to specified combinations of scanned

occupational titles and industry or education information. The computer-assisted approach was then used only for the remaining occupations. Statistics from the automatic coding process suggest that over half of all individuals with occupation were coded automatically. While this process was much faster than computer-assisted or manual coding, it could have resulted in some loss of quality, as information recorded in other fields was not taken into account. In addition, as noted above, the absence of the question on tasks and duties could mean that the quality of the data on occupation is not as good as in 1996.

Statistics South Africa uses the South African Standard Classification of Occupations (SASCO). This classification is very similar to the International Standard Classification of Occupations of 1998 (ISCO-1998), but further disaggregates ISCO-1998's four-digit occupation codes into five-digit codes to allow more detailed analysis of the type of work done. When coding for Census 1996 commenced, the intention was to code to the full five digits of SASCO. This was done for the first few weeks of coding. However, the fully automated coding system coded only to the three-digit level. Because such a large proportion of occupations were coded automatically, Statistics South Africa decided to do all subsequent coding only to three digits.

The consequence of this decision is that some important occupations cannot be distinguished. The most significant of the occupations which are not distinguishable at the three-digit level is domestic workers in private households, who are grouped together with cleaners employed in factories, offices and elsewhere. Fortunately in this case the domestic workers in private households can be distinguished from other cleaners by combining the industry and occupation codes. However, this cannot be done for other occupational groups. Three-digit coding does not, for example, distinguish between accountants and personnel consultants; between different types of engineering technicians and draughtspersons; between travel consultants and estate agents; between athletes and TV announcers; and between undertakers and hairdressers.

On completion of the scanning and data capturing phases, there is a further procedure of editing. For Census 2001, Statistics South Africa was assisted by the US Bureau of the Census which helped establish a sophisticated automatic editing system. This system finds and corrects apparent inconsistencies according to rules which are designed to reflect the most likely 'correct' data. Where there are data gaps, the editing system uses imputation from a 'hot deck' where the likely correct data are not able to be derived from existing

variables for that record by rules of logic. The 'hot deck' approach involves assigning the value for a particular variable in a record from the value for that variable in a record which corresponds to an individual with a similar demographic profile. This approach has been adopted in a number of other countries, often with the assistance of the US Bureau of the Census. It is the first time it is being used by Statistics South Africa (For full details see Statistics South Africa 2003).

The automated approach has a number of advantages. While it involves an enormous amount of work to set up, once established it provides a relatively fast and efficient way of cleaning and 'correcting' data that treats similar cases in a consistent way. It also provides the possibility of comparing the patterns produced by the raw and edited data and modifying edits until the changes achieved are deemed the optimal ones possible.

One danger is that consistency does not necessarily mean that the value allocated is correct. It simply means that the same value will be allocated for similar cases, and that the value will reflect the assumptions of the people who compiled the programme. A particular danger with imputation from a 'hot deck' is that if the number of cases for which this method is used exceeds a very small percentage, one is effectively creating a data set. Monitoring of the imputation process revealed that the percentage of cases imputed from the 'hot deck' was, in fact, far from negligible in respect of some of the employment-related variables.

The imputation rate could be significantly higher than the average for some smaller geographical areas, as missing data are likely to be clustered as a result of the weakness of particular enumerators. Because the data set appears to contain full information for each individual, many users may not recognise that a large proportion of the information for a particular area was not obtained from the household but, instead, is based on hypotheses about likely patterns. Statistics South Africa will release two different 10 per cent samples of the data – one with, and one without, imputation. Researchers will thus be able to investigate the impact of imputation.

The electronic document imaging solution being used for Census 2001 was intended to save time in comparison to the traditional method used in 1996. Unfortunately, Statistics South Africa encountered some teething problems with the use of the technology. Shortly before enumeration commenced, Statistics South Africa became aware that a particular quality of paper and printing was required for the questionnaires if scanning was to be used. The

short time available meant that South African firms were not able to supply the number of printed questionnaires required. This problem was solved – although at some cost – by getting questionnaires printed in the US.

Once scanning started, there were further problems with the technology. These included the size of the different data sets, difficulties with the complexity of the process, initial lack of documentation, system down-time problems, and scanning problems. The problems were eventually solved, but caused delays which meant that the technology did not result in the time savings initially envisaged.

Less easy – if not impossible – to correct were errors introduced by the scanning process. In particular, analysis of the data suggests that the scanning process introduced a significant number of errors where, for example, the scanner did not distinguish correctly between 7s, 9s and 1s, 8s and 3s, and other combinations. In a few cases where the combination of codes for a particular household or individual were either impossible or very unlikely, the scanning errors could be corrected through rules of logic. In most cases, however, this was not possible as the patterns were possible and therefore not picked up.

It is clear that Statistics South Africa has learned a lot through this process. Some of the most important lessons relate to questionnaire design and layout. Careful consideration should be given to the way that the questions are constructed as well as to the possible answers that will be provided. It is also important to continue testing the technology between the censuses to make sure that all systems are working properly. Another very important aspect is that of interdependencies of processes during data conversion. Attention needs to be given to all the processes and how these influence each other.

Post-enumeration survey

In both 1996 and 2001, the census was followed by a post-enumeration survey (PES). It is accepted all over the world that some households and individuals will be missed in an exercise the magnitude of a census. The PES is intended to provide an estimate of the extent and nature of this problem so that the count can be adjusted accordingly. The PES for 2001 will also provide an estimate of the extent to which there are errors in the content i.e. in responses to questions about characteristics of individuals.

Comparison of the PESs of 1996 and 2001 warrants a separate paper as these are complex exercises in their own right. The following table draws a quick comparison between the operations of the PES for Census 1996 and

Table 1 Comparison of Post-enumeration Surveys, Census 1996 and Census 2001

	Census 1996	*Census 2001*
Time frame	Immediately after the completion of census enumeration	Immediately after the completion of census enumeration
Actual date	15 to 24 November 1996	7 November to 7 December 2001
Reference date	*Not applicable*	*Night of 6–7 November*
Sample: Size	800 EAs	600 EAs
Frame	Stratified by Province	Stratified by Province
EA Classification	Formal urban, informal urban, tribal, commercial farms or other non urban.	Formal urban, informal urban, tribal, commercial farms or other non urban.
Method	Independent systematic sample within each stratum.	Independent systematic sample within each stratum.
Field Workers	*Senior census field workers* *1600 interviewers (2 per EA)* *200 supervisors* *50 regional managers*	*Household survey staff* *649 interviewers* *150 supervisors* *9 regional managers* *9 provincial survey managers*
Conducted	Independent of census	Independent of census
Listing	Redone based on census boundaries	Redone based on census boundaries
Exclusion	*Institutions, hostels – attempted but not achieved, homeless* *Empty & nearly empty EAs*	*Institutions, student residence, tourist hotels/motels/inns, and homeless and vacant EAs*
Coverage in EA	*The whole EA was enumerated*	*Whole EA except for mixed EAs, where institutions were excluded.*
Questionnaire	*Brief, covering basic demographic information, whether the household was visited and all counted, opinion towards census.*	*Brief, covering basic demographic information, whether the household members were counted.*
Processing	*Captured by external organisation*	*Internal capturing by independent team. Captured manually after problems with scanning process.*
Matching	*Matching EAs* *Matching households* *Matching persons*	*Matching EAs* *Matching stickers* *Matching households* *Matching persons* *Reconciliation visits for unresolved cases*
Calculation method	*PES data used to calculate an adjustment factor* *Adjustment factor is then applied to census count* *Undercount rate calculated*	*Dual system estimation (matching of records from two sources – the PES and Census – which are independent)*

Italics indicate differences between Census 1996 and Census 2001

Census 2001. The italicised sections indicate where there was a difference between the two exercises.

The table reveals that the two PESs were similar in most respects. However, a number of changes were introduced in an attempt to improve the quality of the PES. For example, the reduction in the number of fieldworkers and the employment of regular household survey staff was intended to ensure better quality control. On the other hand, the strategy of employing two enumerators per EA was not adopted in 2001. This strategy was used in 1996 to speed up the process so as to minimise the time between enumeration and PES and so avoid errors due to loss of memory of respondents. The table reveals that in 2001, the PES took significantly longer to complete than in 1996. This could have contributed to difficulties experienced in matching households and individuals enumerated in the census and PES.

The reconciliation visits in respect of non-matched cases was another innovation in 2001. In 1996, one of the weaknesses of the PES was that 22 per cent of households and persons were given "unresolved" status in terms of whether they could be matched with census records. This necessitated the use of imputation of status and could have resulted in an underestimate of the undercount. In 2001, the PES included a reconciliation stage where households which had unresolved status were revisited in an attempt to determine their matching status.

In terms of calculation, the 1996 method assumed that the PES count was correct, and adjusted the census count accordingly by an "undercount rate". The relevant formulae were simple, and involved very few variables. The adjustment factor was set equal to the number of people in the PES in the scope of the census divided by the number of people in the PES counted in the census. This adjustment factor was then applied to the census count to obtain the final population estimate. The 1996 approach thus implicitly assumed that the PES count was more accurate than the census count.

Unlike the 1996 PES, the method used in 2001 did not assume that either the PES or census was correct, or 'better' than the other. For example, it acknowledged that the census might have correctly counted some households which were missed in the PES. To obtain two independent counts, the 2001 PES introduced the notion of a second enumeration date, the night of 6–7 November. The PES questionnaire then asked about the whereabouts of each person on both enumeration dates. The final matching phase assigned one of nine different match statuses to each record, as follows:

- Matched
- In PES not in Census, missed in Census
- In PES not in Census, PES erroneous inclusion
- In PES not in Census, PES insufficient information
- In PES not in Census, in-mover
- In PES not in Census, born after Census
- In Census not in PES, correctly enumerated in Census, missed in PES
- In Census not in PES, Census erroneous inclusion
- In Census not in PES, Census insufficient information

These nine categories made up two separate samples. The P (population) sample consisted of all persons enumerated in the PES. Six of the options above were possible for the P sample. The E sample (enumeration) sample consisted of all persons enumerated in the Census. Four of the options above were possible for the E sample.

In terms of calculation of the final estimates, the 2001 method assumed that the 'true' population was made up of:

- The matched population;
- The population included in the Census and missed in the PES;
- The population included in the PES and missed in the Census; and
- The population missed in both the Census and the PES

All but the fourth element could be obtained through direct observation, through the matching process. The fourth was derived mathematically, on the assumption of independence of the two counts.

The "undercount" estimate described above focuses on whether an individual was counted or not. There is also the further question as to whether the individual's characteristics were correctly recorded. The PES 2001 provides estimates of this "content error" as well as the undercount estimate. For each of the demographic variables on the questionnaire (sex, age, relationship to head, marital status, population group, and language), the PES provides four measures of response variance. As with the undercount, the PES does not assume that either the census or PES is more accurate. It is thus not able to give any indication of response bias (Statistics South Africa 2003). At the time of writing, the report containing estimates of content error had not yet been released.

Above we pointed to serious problems with matching in 1996. Part of the problem in 1996 was that the census questionnaire required only that the first name or initials of each individual be recorded on the form. This approach was adopted to allay fears about confidentiality. However, it caused problems in

the PES when trying to confirm whether or not the PES and Census had recorded the same individual. Census 2001 avoided this source of uncertainty by requiring that the full name and surname of each person be recorded on the census form.

The use of barcode stickers in 2001 was a further strategy used to assist with matching. In 1996, some dwellings had been identified on the form by the surname of the household head. This caused problems where household heads in a given EA had common surnames. In other cases enumerators used numbering systems such as the numbering on RDP houses to identify dwellings. Problems arose here where there were multiple numbering systems, or where new numbers had been introduced between the Census and PES dates.

For Census 2001, each questionnaire had two detachable barcode stickers on its first page. These were intended to facilitate exact matching of households enumerated during the census and PES. After the census interview was completed, the enumerator detached one barcode sticker and asked the household to attach it to a door, gate, wall, or any other suitable place. The second sticker was handed to the household to keep for at least two months. When the PES enumerator arrived and interviewed the household, they would ask for the second sticker and attach it to the PES questionnaire. If the sticker was lost, the enumerator would write the barcode number on the appropriate space on the questionnaire. The sticker approach worked well, and helped a lot with the matching of census and PES questionnaires.

On the negative side, the timing of the 2001 PES overlapped to some extent with the census proper as the enumeration period for the census was extended for up to three weeks in some areas when it was recognised that enumerators had not been able to complete their tasks in the time originally allocated. In some cases, PES fieldworkers had visits planned to EAs before these EAs had been fully enumerated. This constituted a serious problem as the main question in the PES is whether the household has been enumerated or not. In cases where enumeration for the census was still being conducted, PES enumerators waited until the enumeration in the area was completed before they started.

In addition, like the census proper, the PES faced significant technological problems. The PES questionnaire was designed to be used with scanning to determine matching, and with computer-assisted input of the characteristics of individuals. Unfortunately, the problems with the technology meant that both these processes had to be done manually. This entailed significant delays and some adjustments in processing methodology to provide the necessary checks.

The PES approach adopted for Census 2001 was more sophisticated than the one used for Census 1996. The PES for 2001 revealed a much higher undercount than for Census 1996 – about 17 per cent compared to 10.7 per cent for the previous census. This level of undercount is something that Statistics South Africa will need to give serious attention so as to avoid a recurrence in future censuses.

Concluding remarks

The jury is still out as to whether Census 2001 was better than Census 1996. In truth, the verdict is likely to be that it was better in some respects, but not in others. The key results were released in July 2003, while the full data sets were made available to the public at the end of 2003. The real litmus test will come in the analysis of the full data sets by the academic community and subject matter specialists.

The delays in the release of the results and data already reflects a weakness as Statistics South Africa had initially hoped to release results within a year of completing enumeration i.e. by October 2002. Unfortunately, a range of unforeseen problems caused delays at different stages in the process. An important learning, then, is that the introduction of sophisticated technology does not necessarily immediately solve timing and other challenges, especially the first time the technology is used.

This paper has pointed to the ways in which Statistics South Africa attempted to learn from its experience in both Census 1996 and the pilot for 2001 so as to improve the quality of its operations and output. The main purpose of the article is to make readers aware of what was done, so that they can better understand and evaluate the census results, and use the data with greater insight. The paper has pointed to both weaknesses and strengths in the changes, and in how they were implemented. Ongoing operations and the final results will no doubt provide further lessons. These lessons must, in turn, be used in attempts to improve the quality of future censuses. Statistics South Africa's own desire to learn from its experience is revealed by the comprehensive documentation that was produced on the different stages and processes involved in Census 2001. Some of the sources listed below are public documents, freely available on the web. Others were written more for internal usage, but can be obtained from Statistics South Africa on request.

References

Adegboyega, O. 2001. "Deliberation of Gaborone Workshop on SADC 2000 Census Round Questionnaire." Report back paper presented to Statistics South Africa, November 2001.

Central Statistical Service. 1996a. *Manual for the Enumeration of Institutions*. Fieldwork Manual. Pretoria: Central Statistical Service.

Central Statistical Service. 1996b. *Enumerator's Manual*. Fieldwork Manual. Pretoria: Central Statistical Service.

Central Statistical Service. 1996c. *Chief Enumerator's Manual*. Fieldwork Manual. Pretoria: Central Statistical Service.

Central Statistical Service. 1996d. *Controller's Manual*. Fieldwork Manual. Pretoria: Central Statistical Service.

Central Statistical Service. 1998a. *The count and how It was done. Pretoria*. Report No. 03-01-17 (1996). Pretoria: Central Statistical Service.

Central Statistical Service. 1998b. *Population census, 1996 Definitions*. Census 1996 Metadata. Pretoria: Central Statistical Service.

North, H. 2002. "GIS process poster", Poster presented at the Statistics Day Celebrations, Durban, November 2002.

Procon. 2001. *Census 2001 Training for the Data Processing Solution*. Internal Training Manual. Pretoria: Procon.

Statistics South Africa. 2001b. *Census 2001 Enumerator's Manual*. Fieldwork Manual. Pretoria: Statistics South Africa.

Statistics South Africa. 2001c. *Census 2001 Listing Manual*. Fieldwork Manual. Pretoria: Statistics South Africa.

Statistics South Africa. 2001d. *Census 2001 Fieldwork Co-ordinator's Manual*. Fieldwork Manual. Pretoria: Statistics South Africa.

Statistics South Africa. 2001e. *Census 2001 Supervisor's Manual*. Fieldwork Manual. Pretoria: Statistics South Africa.

Statistics South Africa. 2001f. *Census 2001 Map reading Manual*. Fieldwork Manual. Pretoria: Statistics South Africa.

Statistics South Africa. 2002: *Post enumeration survey: Preliminary report on methodology*. Pretoria: Statistics South Africa.

Statistics South Africa. 2003. *Census 2001 Computer Editing Specifications*. Pretoria: Statistics South Africa.