

## ATTACHMENT

### THE NEW CONSTRUCTED AND EXTENDED CENSUS 2001 EA SAMPLING FRAME AS WELL AS A BENCHMARKED VERSION TO THE STATSSA MIDYEAR POPULATION ESTIMATES 2005

(Constructed and compiled by D.J. Stoker, 5 June 2006)

The new constructed Census2001 EA sampling frame was developed due to the necessity to have available an EA sampling frame which can be used for the design of area based samples and master samples. This construction was necessitated by the fact that StatsSA has decided not to release census 2001 EA based data. The smallest area based data released by StatsSA are the Small Area Layer (SAL) data with their spatial data. Where there are 80787 EA's used in the 2001 census, the SAL data consists of 56255 small areas. These small areas cannot be related to the 2001 census EA's apart from spatially overlaying the set of SAL's on the set of EA's. (Note that although the SAL data contain Main Place and Sub Place names and codes, new Sub Place names and codes were introduced, especially in rural areas. This makes a one-to-one matching based on Sub Place codes impossible – see below.). The GIS section of the HSRC has done this spatially overlaying of the SAL's on the set of the census 2001 EA's and use have been made of this spatially overlaying to establish a relationship between the 2001 EA's and the SAL's.

The new census 2001 EA sampling frame was constructed as follows:- The 1996 census data were released at the 1996 EA level. In 1996 there was about 94000 EA's. By using transformation information on the relationship between the 1996 set of EA's (about 94000 EA's) and the 2001 set of EA's (80787 EA's), the 1996 released EA data were "carried over" to (or superimposed on) the 2001 set of EA's. By making use of the SAL data, Supercross 2001 and other data sources published by StatsSA, such as "Statistics in Brief 2001", the 1996 census data "carried over" to the 2001 set of EA's were then benchmarked at the Small Place (or Sub Place) level to meet the available StatsSA data at this level. This was an enormous exercise. As benchmark variables were used:

- Population size
- Number of households
- Number of males
- Number of females
- Number of Blacks (Africans)
- Number of Coloureds
- Number of Asians/Indians
- Number of Whites
- Number of persons per 5-year age group
- Number of persons per highest level of education for all education categories
- Number of persons per home language for all 11 SA home languages
- Number of heads of households separately for males and females
- Number of households per dwelling type for all dwelling types

It is not claimed that this information is correct at the EA level, but only at the Small (Sub) Place level.. Note that there may be or even will be differences in the variable values at Small Places forming part of 70 Main Places for which new Sub Places and Sub Place codes were introduced by StatsSA. For these cases the data will be correct at the Main Place level. In respect of the variables: highest education level, home language, head of household and dwelling type the data were benchmarked only at the sub place level. Note further that benchmarking has not been done for combinations of the benchmark variables, such as gender by race. This is possible but will cause an enormous amount of additional work.

The above data were also benchmarked to the midyear 2005 estimates as released by StatsSA. This benchmarking was only done in respect of those variables for which such marginal values were available, viz. province, gender, 5-year age group and race. These data are stored in a separate file.

#### List of variables on sampling frame file

EA\_number= number of enumeration area  
EA\_type\_desc= description of EA type  
EAtype= numerical variable for EA type  
EA\_geography\_type= geography type of EA  
Geotype= numerical variable for geography type (1 to 4)  
Prov\_name= Province name  
province= numerical variable for province (1 to 9)  
mp\_code= main place code  
mp\_name= main place name  
mp\_type= main place type  
sp\_code= sub place code  
sp\_name= sub place name  
sp\_type= sub place type  
region\_code= region code  
region\_name= region name  
MD\_code= magisterial district code  
MD= magisterial district name  
DC\_code= district council code  
DC\_name= district council name  
munic\_code= municipality code  
munic\_name= municipality name  
totalpop= total population size  
total\_HH= total number of households  
males= number of males  
females= number of females  
blacks= number of blacks  
coloureds= number of coloureds  
asians= number of asians  
whites= number of whites  
age0\_4= number of persons in the age group 0 to 4 years

age5\_9= number of persons in the age group 5 to 9 years  
age10\_14= number of persons in the age group 10 to 14 years  
age15\_19= number of persons in the age group 15 to 19 years  
age20\_24= number of persons in the age group 20 to 24 years  
age25\_29= number of persons in the age group 25 to 29 years  
age30\_34= number of persons in the age group 30 to 34 years  
age35\_39= number of persons in the age group 35 to 39 years  
age40\_44= number of persons in the age group 40 to 44 years  
age45\_49= number of persons in the age group 45 to 49 years  
age50\_54= number of persons in the age group 50 to 54 years  
age55\_59= number of persons in the age group 55 to 59 years  
age60\_64= number of persons in the age group 60 to 64 years  
age65\_69= number of persons in the age group 65 to 69 years  
age70\_74= number of persons in the age group 70 to 74 years  
age75\_79= number of persons in the age group 75 to 79 years  
age80\_84= number of persons in the age group 80 to 84 years  
age85over= number of persons 85 years and over  
aged80over= number of persons 80 years and over  
grade\_1= number of persons with grade\_1  
grade\_2= number of persons with grade\_2  
grade\_3= number of persons with grade\_3  
grade\_4= number of persons with grade\_4  
grade\_5= number of persons with grade\_5  
grade\_6= number of persons with grade\_6  
grade\_7= number of persons with grade\_7  
grade\_8= number of persons with grade\_8  
grade\_9= number of persons with grade\_9  
grade\_10= number of persons with grade\_10  
grade\_11= number of persons with grade\_11  
grade\_12= number of persons with grade\_12  
Certdip\_less\_Grade\_12= number of persons without Gr12 with a certificate or a diploma  
Cert\_with\_Grade\_12= number of persons with Gr12 and a certificate  
Dipl\_with\_grade\_12= number of persons with Gr12 and a diploma  
B\_degree= number of persons with Bach. degree;  
B\_degree\_dip\_hons= number of persons with Bach. degree with diploma or with Honn degree  
M\_D\_degree= number of persons with M\_D\_degree  
No\_school= number of persons with no schooling (not applicable)  
NA= Not applicable  
afrikaans= number of persons with Afrikaans as home language  
english= number of persons with English as home language  
Ndebele= number of persons with Isindebele as home language  
Xhosa= number of persons with Isixhosa as home language  
zulu= number of persons with Isizulu as home language  
sepedi= number of persons with Sepedi as home language

sesotho= number of persons with Sesotho as home language  
 setswana= number of persons with Setswana as home language  
 siswati= number of persons with Siswati as home language  
 venda= number of persons with Tshivenda as home language  
 tsonga= number of persons with Xitsonga as home language  
 Lang\_oth= number of persons with other language as home language  
 H\_separa= number of houses on a separate stand or yard  
 trad\_dwl= number of traditional dwellings or huts  
 flat= number of flats in block of flats  
 town\_hse= number of town houses or clusters or semi-detached houses  
 flatyard= number of houses or rooms or flats in backyard  
 shckyard= number of informal dwellings or shacks in backyard  
 shckelse= number of informal dwellings or shacks NOT in backyard  
 shareprp= number of flatlets or rooms shared property (not in backyard)  
 Dwl\_other= number of persons in other dwellings (special institution such as  
 caravan/tent/private shipboat)

**Remarks:**

It is important for the user of any of these data sets to be aware of, amongst others, the following proviso's (the discussion below concentrates on census 2001):-

1. It is not claimed that the data per EA are identical with the unreleased EA data resulting from StatsSA census 2001 data files. There may thus be considerable differences in the values of variables per EA. In the drawing of a complex sample the variables mostly used as measure of size in the drawing of the primary sampling units (i.e. EA's) are the number of households or the number of persons within an age interval (such as between 15 and 65). A sampling frame needs not be 100% correct in respect of the chosen measure of size values to be used in the drawing of the sample provided that corrections or adjustments are made afterwards based on information obtained during the fieldwork exercise. However, a sampling frame should be complete with regard to the sampling units (e.g. EA's) that it contains. Note that benchmarking of the adjusted sample record weights in respect of available (i.e. released) information from census 2001 is vital to reduce any possible bias resulting from the use of the constructed EA sampling frame (apart from reducing bias resulting from other sources such as the fieldwork "errors").
2. In a very few EA's the estimated number of households was found to be larger than the estimated number of persons, which is clearly not possible. This was caused by the benchmarking process.
3. It is important to realize that the EA descriptive information given on the sampling frame can be incorrect for some EA's. The information is, however, exactly the same as the information obtained during the demarcation phase in preparation for the 2001 census. If an EA descriptive variable (such as geography type or EA type) was used as an explicit stratification variable in the drawing of the sample of EA's, incorrect information about the geography type of an EA found during the fieldwork can be adjusted for by re-calculating the sampling weights of the drawn EA's afterwards. (For example, the variable EA type

- description indicates in many cases incorrectly that EA's were vacant.) The above remark is not applicable to variables used as implicit stratification variables.
4. It is well known that EA's were "missed" in the 2001 census enumeration phase. This is one reason why the constructed 2001 EA sampling frame could have data for "empty" EA's in census 2001. It is, on the other hand, also possible that a "non-empty" EA in the 2001 census could be "empty" in the constructed EA sampling frame due to the fact that it was "empty" in the 1996 census (if, for example, a new development took place between 1996 and 2001). It is thus clear that it cannot be claimed that the constructed EA 2001 sampling frame is complete iro of its non-empty EA's. But, apart from the fact that EA's were "missed" in the enumeration stage of census 2001, new developments have taken place since 2001, so that the sampling frame directly based on the unreleased census 2001 EA information is also incomplete iro of the "non-empty" EA's that it contains. There is, furthermore, an erratic movement of many squatter areas. It is again emphasized that fieldwork teams should check the correctness of the descriptive information of EA's drawn into the sample. Any EA's drawn as part of the sample which was found "empty" during the fieldwork may "legally" be replaced by drawing another "similar" EA, but again this will require a re-calculation of the sampling weights of the EA's afterwards. It must be emphasized to fieldwork teams to report all relevant information as indicated above to their office.
  5. The constructed EA sampling frame contains only 80780 instead of 80787 EA's. Seven EA's in Mouille Point in Cape Town had to be excluded due to the total lack of any usable information.

### **Final Remark**

It will be appreciated if the user of any of the new constructed EA sampling frames report to the constructor/compiler of these sampling frames any inaccuracies or errors found in the use of the sampling frames.