APPENDIX B: Processing Population Register Data

Data source

Copies of the log files for transactions entered on the population register regarding deaths for the period 1997 to date have been obtained through different channels. Different sources were necessary since the Department of Home Affairs only retains their log tapes for a 12-month period, and an extraction of deaths from the full Population Register was not practical.

i) The first set of data was obtained from Stats SA who had been provided with a 12-month set of the deaths from the log file for the period 1997/98.

ii) The second set of data for the period 1998/99 was obtained from the Department of Health who had also obtained a copy of the DHA log files. The Department of Health cleaned their data files and removed all duplicate records.

iii) Data from April 1999 onwards has been obtained directly from the Department of Home Affairs.

Multiple records

Multiples were identified in the combined data and any that had the identical ID number were removed, keeping the latest record only. Multiples arise from the fact that when any administrative amendments regarding the death registration are made to the record on the Population Register, it is logged.

Age

The ID number has 13 characters. The first 6 digits reflect the date of birth. In the first two data sets, the age is calculated by subtracting the date of birth component of the ID variable from the date of death. A few cases have non-standard ID numbers, thus the date of birth details could not be taken from the ID number and age could not be calculated for them.

The data sets obtained directly from the Department of Home Affairs also contain the actual date of birth of the deceased. Where in some cases, only the year of birth was available, age was calculated by subtracting the year of birth from the year of death.

The ID number does not indicate the century of birth. However for all deaths, with a date of birth of the deceased being 2000, we have a corresponding date of birth as a separate field.

Sex

Digits 7 to 10 of the ID number indicate the sex of the person (combined with a serial number).

Codes 0000 to 4999 represent females and 5000 to 9999 males.

Cause of death

The text in the cause of death field was analysed and the terms that occurred relatively frequently (more than 10 times in a year) were examined and identified as external or natural causes. Those terms that occurred infrequently were taken as natural causes without examination. The majority of the external causes were recorded as "non-natural" or "Onnatuurlik" but there was a range of specific causes such as "MVA", "head injuries", "wond", "wound", "gun", "shot", "murder", "trauma", etc. Cause of death was either in Afrikaans or English, while the words were spelled in various ways. This was all taken into account when assessing the external causes of death.

Cleaning the data

The data for June 1999 - May 2000 contained some illegal characters for some variables. In these cases, the complete record was examined for obvious problems such as a consistent column shift and was then realigned. In 4 cases the sex could not be identified due to illegal characters in the ID number. When the illegal characters appeared in the date of birth variable, it was possible to correct it by using the date of birth component of the ID variable and vice versa.

The number of exclusions from each data source of the data is summarized in Table B1.

APPENDIX B: Processing Population Register Data

Table B1: The number of records excluded for analysis according to source

Source	Period	Total Records	Multiples excluded	Records Cleaned	Sex unknown	Age unknown	Record for analysis
i)	June 1997 - July 1988	284 391	4 973	279 418	-	2 061	277 357
ii)	June1998 - May 1999	314 881	-	314 881	-	1 895	312 986
iii)	June 1999 - May 2000	360 217	8 461	351 756	4	-	351 752
	June 2000 - Sept 2000	131 038	1 362	129 676	-	-	129 676

Data included in analysis

The three sources for data from Home Affairs indicted above were pooled, and three 12-month periods extracted. Since deaths can take up to two months to be registered we decided to start the first 12-month period in August 1997. This means that the first two periods overlap for the month of July 1998. It should also be noted that while the final period is until June 2000, it includes deaths which occurred in that month but appear on the log tapes up until September 2000.

Only deaths in the age range 15-99 are included in the analysis. The numbers of deaths in each period are shown in Table B2.

Table B2: Composition of annual data sets 1997 to 2000

Period	Total All ages	Male 15 - 99	Female 15 - 99	Total 15 - 99
Aug 1997 - July 1998	278 311	150 878	119 085	269 963
July 1998 - June 1999	306 741	164 873	132 470	297 343
July 1999 - June 2000	343 535	179 109	153 514	332 623