Investigating old age exaggeration in South African population and survey data using near extinct generations methods

Ronald Richman[1,2] and Rob Dorrington[2]
[1] AIG South Africa, [2]CARE, University of Cape Town, South Africa,

Abstract
In order to investigate possible age exaggeration in South African census and survey data, we adapt the near extinct generations method of Das Gupta (1991) to the South African environment. We propose a smoothing mechanism based on the Gompertz curve which is capable of dealing with age exaggeration and heaping in the death data. Since the method relies on completely reported death data, we correct the death data for under-reporting using Death Distribution methods before the application of the smoothed near extinct generations method. We find that significant age exaggeration is present in the Census data of 2011 from the age of 95 and provide a confidence band around our estimates. We also find that date of birth digit preference occurs in both the population and death data, which highlights the need for smoothing of the death data prior to using them to correct population estimates.

Background
It is well known that population data, particularly in developing countries, often suffer from the inaccurate recording of the age of the elderly see for example Preston, Elo, Rosenwaike *et al.* (1996)- often manifesting in the aggregate data as too many being recorded at older ages (Preston, Elo and Stewart 1999). Inaccurate recording of age could be estimated directly, by examining matched census and birth records (Hill, Preston and Rosenwaike 2000; Preston, Elo, Rosenwaike *et al.* 1996), but, even if the data are available to do this, a direct investigation is costly and suffers from sampling error. An alternative is to use demographic methods to reconstruct the old age population and indirectly measure the extent of age misreporting. A key technique used to reconstruct the old age population is the method of extinct generations and its extension to nearly extinct generations ("NEG")(Andreev 1999; Andreev 2004; Das Gupta 1991; Thatcher, Kannisto and Andreev 2002), which rely on death data, which are usually more accurate than census and survey data.

Whilst the method of extinct generations has been applied quite widely to developed-country populations (Coale and Caselli 1990; Coale and Kisker 1990; Thatcher 1992; Thatcher, Kannisto and Andreev 2002), the application of the method to developing country data is complicated by death data that are incompletely recorded and that may also suffer from age misreporting, in the form of digit preference and age exaggeration. In addition, whilst the population estimates produced by these methods suffer from the uncertainties inherent in all projection methods, the quantification of this uncertainty via statistical methods is still a relatively under-researched topic. Our study builds on the results of Machemedze (2009).

Aims, data and methods
This research shows that the NEG methods can be adapted to the challenges presented by developing country data, when these data have been corrected for under reporting of deaths through death distribution methods.

The death data are corrected for under-registration by applying Death Distribution methods ("DDMs") (Bennett and Horiuchi 1981; Hill 1987) to the data at both a national level and to ethnic

sub-populations. The estimates of completeness produced by these methods for intercensal periods are interpolated by fitting a log-logistic curve to the estimates for several inter-censal periods to provide correction factors for the death data for individual years, leading to the estimates of the true number of deaths at each age for each year from 1996 to 2013.

The NEG methods are adapted to deal with exaggeration of age in the death data used to fit the projection model by applying smoothing during the calculation of the projected deaths. This smoothing removes the effect of age heaping on the future deaths projected from the initial death data but does not correct the initial deaths (that are projected) for age heaping, allowing us to investigate whether the observed age heaping is consistent in both the population and death data. We focus on the Das Gupta (1991) method which predicts deaths to a cohort aged $x+1$ in the year $t+1$ using the ratio $\frac{D_{x+1,t+1}}{D_{x,t}}$ calculated using data from previous years, known as a "cohort ratio".

Building on the principle in Thatcher (1990), who provides formulae for the curve of deaths based on the Gompertz curve of mortality, we simplify the cohort ratio on the assumption that mortality follows a Gompertz curve. We are therefore able to find a parametric form for this ratio which smooths out fluctuations resulting from misreported data.

This adapted method is then applied to the death data corrected for under-registration at a national level for the years 1997-2013 to reconstruct the old age population in the years 1996, 2001, 2007 and 2011. The reconstructed population is compared to the enumerated population to assess the extent of age exaggeration in the population data. In addition, we construct confidence intervals around the best estimate of the reconstructed population are constructed by applying a bootstrap procedure to the death data which allows the quantification of the uncertainty in the estimates.

Results and anticipated conclusions

Completeness of reporting of deaths rose rapidly in the late 1990s to reach over 90% by mid-2000 levelling off at around 94% towards 2010.

We present illustrative results using death data for males for the period 1996-2013 and the censuses of 1996 and 2011. The cohort ratios and the smoothed ratios for 2011 males are presented in Figure 1.

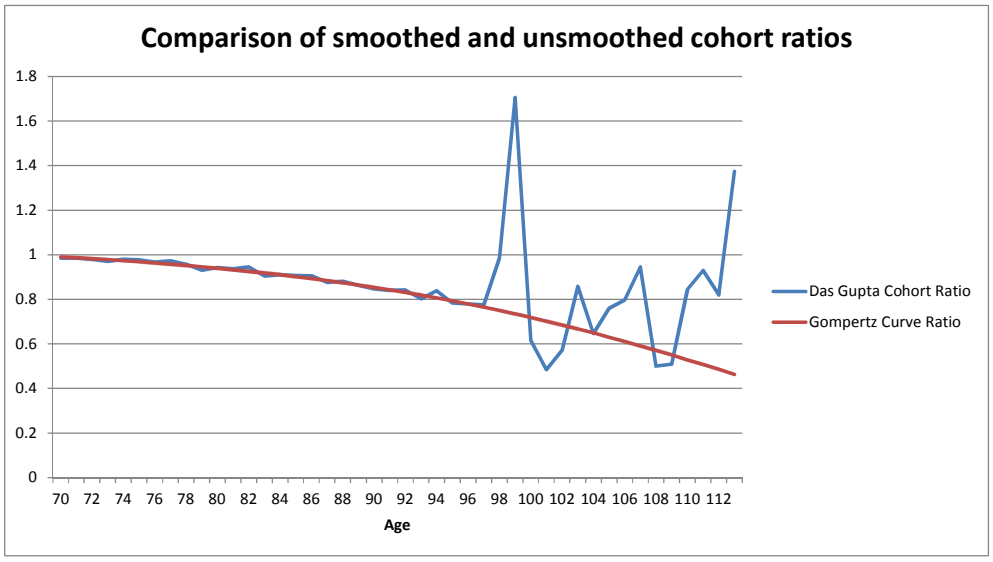

Figure 1: Comparison of smoothed and unsmoothed ratios for males 2011

The smoothing process is seen to remove the very large fluctuations in the ratios at the advanced ages. The smoothed ratios follow the unsmoothed ratios closely for younger ages but trend downwards compared to the unsmoothed ratios at older ages. The use of the unsmoothed ratios would, on average, produce estimates of the population that are larger than the smoothed ratios, since the smoothed ratios lie consistently beneath the unsmoothed ratios.

The reconstructed male population in 2011 is compared to the recorded population in 2011 in Figure 2. The figure shows the numbers (left hand axis) and the ratio of the count to the smoothed numbers (right axis).

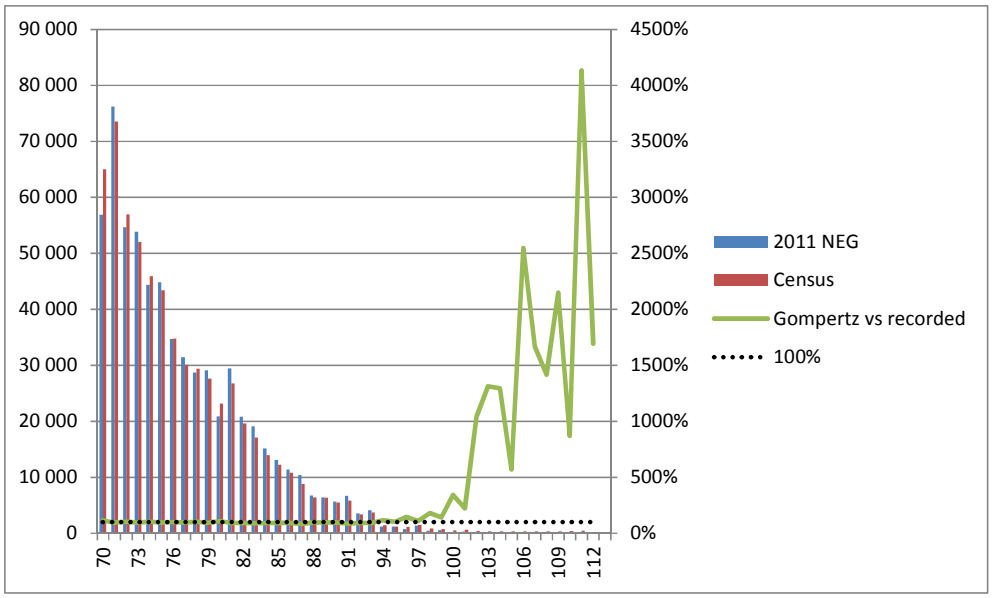

Figure 2: Comparison of reconstructed and recorded population, Males, 2011.

This comparison shows that age exaggeration is most apparent from around age 95, a finding consistent with Machemedze (2009) who investigated age exaggeration in an earlier census. However, the impact of age heaping is not removed in the population re-estimated from death data, which were expected to be more accurate than the population data. The heaping is observed in both the enumerated and re-estimated populations at ages which imply a year of birth ending in a '0' which suggests that approximate years of birth have been used on birth certificates, leading to similar "reverse age heaping" (Myers 1976) in both sets of data. We therefore conclude that the death data should be smoothed before projection (results not shown).

The impact of using the smoothed versus the unsmoothed ratios on the size of the re-estimated population is shown in Figure 3.

A notable "dip" occurs between age 95 and 100, where the Gompertz curve has smoothed out the large fluctuations in the cohort ratios (caused by age heaping) that is shown in Figure 1. The unsmoothed reconstructed population is generally materially higher than the smoothed population, from around age 80. This illustrates that age has been exaggerated in the South African death data for males or that the mortality of South African males does not follow a Gompertz law, or both. Although it has been shown that that mortality in some developing countries (e.g. China (Yi and Vaupel 2006)) does not follow a Gompertz law, South African data to this point have not been complete and accurate enough to decide whether this is also the case in South Africa. However, given that the unsmoothed ratios lie above the smoothed ratios in Figure 1, it is likely that a greater part of the difference between the two sets of reconstructed population estimates is due to age

exaggeration in the death data, which has been corrected in the "smoothed" estimates by the Gompertz curve.

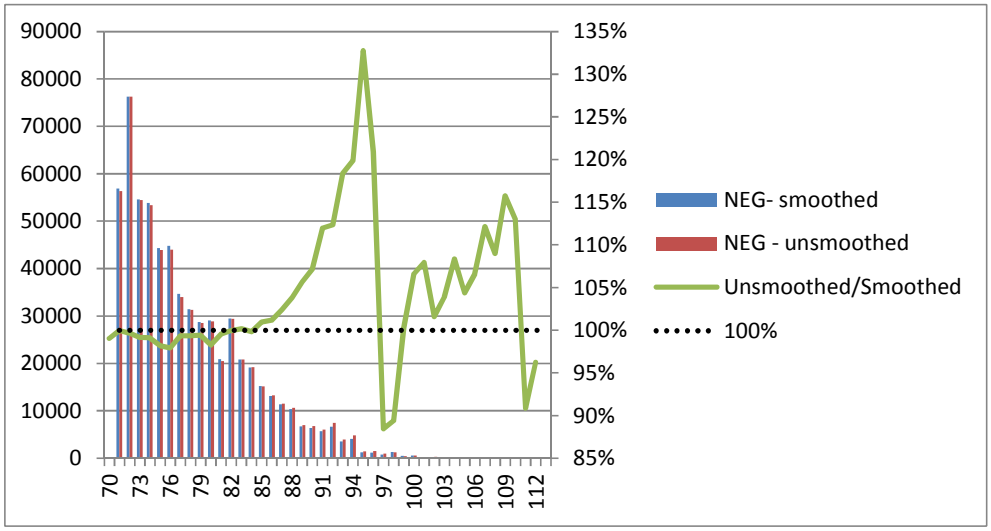

Figure 3: Comparison of two methods of reconstructing the population – smoothed (Gompertz) versus unsmoothed (Das Gupta)

Lastly, in Figure 4 we present the results of a bootstrap procedure applied using the smoothed procedure, for the ages 91-110. We bootstrapped the fitted population 100 times[1] using the residuals of the fitting procedure to resample the death data and to refit the model and then compared the reconstructed population to the enumerated population in 2011. The percentiles of this comparison are illustrated in the figure.

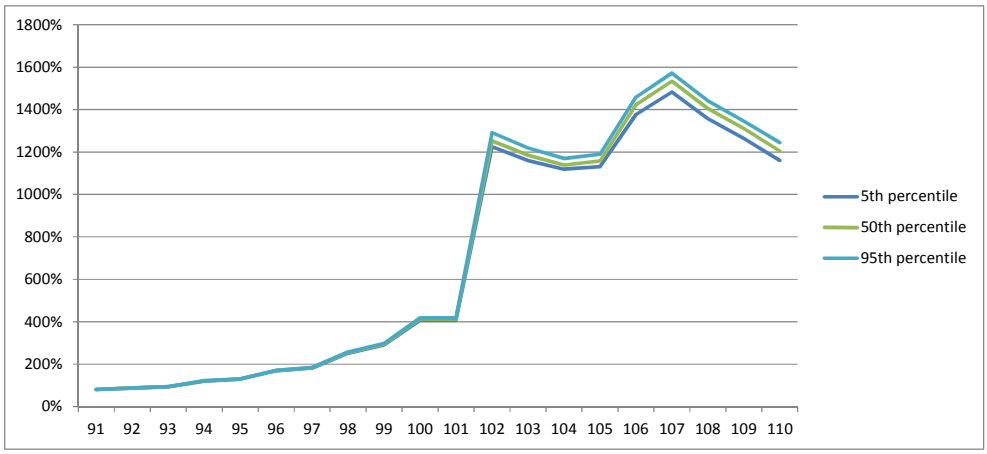

Figure 4: Percentiles of the ratio of the reconstructed population to the enumerated population, Males, 2011

The bootstrap procedure reveals that the projections do not suffer from uncertainty introduced by a poorly fitting model since the variability of the bootstrapped distribution is low over the ages 90-110 (and even more so for ages 70-90, not shown). We are therefore confident that the predictions made to the population are statistically justifiable.

[1] A procedure similar to England and Verrall (1999) was used by calculating the residuals implied by the smoothed method calculated as $\varepsilon_i = \frac{(D_{x,t} - \hat{D}_{x,t}^{smoothed})}{D_{x,t}}$, fitting a normal distribution to the residuals, recalculating the deaths as $\widetilde{D}_{x,t} = \sqrt{\hat{D}_{x,t}^{smoothed}}\,\varepsilon_i + \hat{D}_{x,t}^{smoothed}$ and refitting the smoothed curve to the $\widetilde{D}_{x,t}$.

A second iteration of the fitting procedure will be applied to death data that have been smoothed to account for digit preference. Since each ethnic sub-population exhibits a different pattern of mortality, more accurate projections of the future deaths will be made when fitting the model to each sub-population separately.

References

Andreev, K. F. 1999. "Demographic surfaces: Estimation, Assessment and Presentation, with Application to Danish Mortality." Unpublished thesis, Odense: University of Southern Denmark.

Andreev, Kirill F. 2004. "A method for estimating size of population aged 90 and over with application to the 2000 US census data", *Demogr Res* **11**:235-262.

Bennett, Neil G and Shiro Horiuchi. 1981. "Estimating the completeness of death registration in a closed population", *Population Index*:207-221.

Coale, Ansley J and Graziella Caselli. 1990. "Estimation of the number of persons at advanced ages from the number of deaths at each age in the given year and adjacent years", *Genus*:1-23.

Coale, Ansley J and Ellen E Kisker. 1990. "Defects in data on old-age mortality in the United States: new procedures for calculating mortality schedules and life tables at the highest ages",

Das Gupta, P. 1991. "Reconstruction of the Age Distribution of the Extreme Aged in the 1980 Census by the Method of Extinct Generations," Paper presented at American Statistical Association Proceedings of the Social Statistics Section. 154-159.

Hill, Kenneth. 1987. "Estimating census and death registration completeness," Paper presented at Asian and Pacific population forum/East-West Population Institute, East-West Center. Vol. 1:8.

Hill, Mark E, Samuel H Preston and Ira Rosenwaike. 2000. "Age reporting among white Americans aged 85+: Results of a record linkage study", *Demography* **37**(2):175-186.

Machemedze, Takwanisa. 2009. "Old age mortality in South Africa." Unpublished thesis, University of Cape Town.

Myers, Robert J. 1976. "An instance of reverse heaping of ages", *Demography*:577-580.

Preston, Samuel H, Irma T Elo, Ira Rosenwaike and Mark Hill. 1996. "African-American mortality at older ages: Results of a matching study", *Demography* **33**(2):193-209.

Preston, Samuel H, Irma T Elo and Quincy Stewart. 1999. "Effects of age misreporting on mortality estimates at older ages", *Population Studies* **53**(2):165-177.

Thatcher, A Roger. 1992. "Trends in numbers and mortality at high ages in England and Wales", *Population Studies* **46**(3):411-426.

Thatcher, A Roger, Vaïnö Kannisto and Kirill Andreev. 2002. "The survivor ratio method for estimating numbers at high ages", *Demographic Research* **6**(1):2-15.

Thatcher, AR. 1990. "Some results on the Gompertz and Heligman and Pollard laws of mortality", *Journal of the Institute of Actuaries* **117**(01):135-149.

Yi, Zeng and James W Vaupel. 2006. "Oldest-old mortality in China," in *Human Longevity, Individual Life Duration, and the Growth of the Oldest-Old Population*. Springer, pp. 87-110.