A Methodology for Projecting sub-National Populations Allowing for the Impact of HIV/AIDS and where Data are Limited and Defective

Rob Dorrington and Tom A Moultrie
Centre for Actuarial Research, *University of Cape Town*

Sub-national population projections are complex at the best of times. As the population being modelled represents a potentially small proportion of the national whole, the least understood of the three major demographic forces – migration (local and international) – plays an increasingly significant role, while, in addition, sub-population specific estimates of the other demographic variables may not be available or reliable. This paper describes an approach to projecting sub-national population dynamics in 28 districts of a country (Botswana) whose aggregate population is less than one-fifth that of New York City. Apart from having to ensure that the sum of the regional projections remain consistent with that of the national population, the projection needs to incorporate HIV epidemiological dynamics in each region and work with severely limited and defective census and survey data. In the process, interesting insights into the regional demographic dynamics of the country are developed.

Introduction

The projection of small-area (and sub-national) populations offers several distinct challenges to demographers. First among these is that as the population of the area being projected gets smaller, closer attention needs to be paid to both regional and international migration patterns, levels and trends. The second is to parameterise accurately the other component demographic forces of fertility and mortality. Both of these challenges become more complex when the sum of several small-area projections has to be constrained to reproduce, or closely emulate, that of a larger (e.g. national) aggregate; when one needs take into account complex mortality dynamics associated with generalised AIDS epidemics in each of the constituent subpopulations which could vary in timing and extent; and when the quality of the data to hand to perform the projections is decidedly poor at a national level, and demonstrably worse for the constituent sub-populations.

In this paper, we describe the methodology applied to project the population of each of Botswana's 28 census districts, from 1980 through to 2020, taking into account the impact of HIV/AIDS on the fertility and mortality of the population, as well as allowing for both internal and international migration, and we discuss the results so derived.

Botswana is a small landlocked country in Southern Africa, with a population of some 1.68 million as enumerated in the 2001 census (Central Statistics Office 2004)[1]. A large proportion (around a third in 2001) of the population live in the six main urban centres, although many urban dwellers retain strong links to their rural communities of origin with high level of circular migration which in no small manner accounts for the widespread HIV/AIDS epidemic in the country (in terms of prevalence, one of the highest in the world) and the small urban-rural differential in HIV prevalence. Embedded within this is a system of international migrant labour (largely to South Africa). In the most recent survey conducted in 2005, around a third of women attending antenatal clinics were found to be infected with HIV (Ministry of Health and Child Welfare (Zimbabwe); US Centers of Disease Control and Prevention; UNAIDS 2004). A more widely representative investigation (the Botswana Aids Impact Study, conducted in 2004) indicated that 29.4 per cent of adult women were infected with HIV (National AIDS Co-ordinating Agency (NACA), Central Statistics Office and Partners 2005)[2]. It is in this context that the Centre for Actuarial Research (CARe) at the University of Cape Town, South Africa, undertook (for UNDP and National AIDS Coordinating Agency, Botswana) population projections of the Botswana population, by administrative and census district, through to 2021, making as much use of local data as possible (Dorrington, Moultrie and Daniel 2006). An associated study by a second group of consultants took the results from these projections as inputs into an assessment of the economic implications of the demographic trends projected (Econsult 2006).

The focus of this paper is on applied demography. Consequently, the paper will first describe in greater detail the factors which make the production of sub-national population projections in Africa more challenging than is typically the case. Second, it will describe the method adopted

[1] No Post-Enumeration Survey was conducted, and the enumerated population was not adjusted by the CSO for an acknowledged and evident undercount – particularly of young children and infants. The results from the modelling and population projection exercise undertaken that is the foundation of this paper suggests that the national population in 2001 was some 7 per cent larger, at 1.797 million (Dorrington, Moultrie and Daniel 2006).

[2] Analysis of the potential biases in the two surveys lead to the conclusion that prevalence of adult women in Botswana (15-49) in 2004 was around 31.5% with district prevalence ranging between 20% and 50%.

for producing the constrained sub-national population projections, including the application of a host of demographic techniques, not least among which are those associated with small-area population projections, and the modifications applied to those methods to meet the challenges. Third, we will present results from our projections showing the successful operation of the method; and finally, we will reflect on the process undergone and the extent to which the lessons learned may be of value to future practitioners of population projections in sub-Saharan Africa.

Sub-national population projections in Southern Africa

The derivation of population projections for small areas is not new in the demographic literature, and the forces that render their results less reliable than those for larger populations are mostly well understood. Smith (2003) suggests four problems that apply particularly to small-area population projections, each of which may bear on the type of model used to project the population. First, administrative boundaries of small areas are more likely to change over time. This certainly happened in Botswana: several of the census districts used in the 2001 census were not demarcated separately in earlier censuses. However, the need to allow accurately for the spread of the HIV epidemic in a population meant that the projections had to start in 1980, and estimates of the population delimited according to the 2001 district boundaries had to be derived for earlier censuses. Second, data are sometimes not available in the desired spatial framework. For this project this was problematic only insofar as it relates to the first problem above – namely that in early years of the projections, assumptions of fertility, mortality and migration had to be made in addition to deciding on a base population for districts that did not exist at the desired date. Third, the reliability of the data at increasingly fine levels of disaggregation becomes ever more questionable, not least simply because the inherent error in the estimates increases as both numerators and denominators in derived rates become smaller. As we describe below, this problem takes on whole new aspects of complexity when dealing with demographic and epidemiological data from (Southern) Africa. And finally; Smith observes that small-area projections are more likely to be affected by 'idiosyncratic factors' – such as the siting of a military base, or the delimitation of a national park – specific to the area that might significantly impact on population dynamics. Of course, such idiosyncrasies are inherently unforeseeable, and hence can be incorporated into population projections only very rarely.

Three variations to the issues identified by Smith present themselves in pursuing small-area population projections of sub-Saharan African countries. The first is that commonly encountered when projecting the population of all constituent areas of a larger area, namely that the aggregate of the individual projections must remain consistent with a projection of the larger area done directly. Typically, this consistency must be ensured by both age and sex. In the exercise undertaken here, additional dimensions must also remain consistent, in particular the estimates of national and aggregated districts HIV prevalence, adult AIDS-related mortality and perinatal AIDS-related mortality.

Second, since both epidemiological rates and aggregate populations are required to remain consistent, this requires that aggregated numbers of people in different states (for example, the numbers of people newly infected each year with HIV) has to remain consistent too between the aggregated and directly projected national populations. The inclusion of separate epidemiological dynamics (in terms of the start date, rate of spread (incidence), and level of prevalence) into the projections for each region presents a second complication, since the projections become increasingly non-linear between regions.

The final complication that must be faced here is one that – on some levels – relates to the quality of the data used as inputs in projections. Where vital registration systems and/or census and survey data are of consistently high quality, this is less of a problem. In Africa, this problem is especially severe (Bradshaw and Timaeus 2006; Hill 1990) With the exception of South Africa, no Southern African country has a vital registration system that is even remotely complete[3]. Births and deaths (particularly those of very young children) are routinely significantly underreported. And, while the frequency of censuses has regularised itself after significant effort on the part of the UN in the 1970s and 1980s, coverage and content errors are still widespread. In many instances, no attempt is even made to assess coverage errors in censuses – even in the few cases where post-enumeration surveys are conducted, these are frequently not written up or used in the assessment of the data collected. In all these respects, Botswana is no exception.

The programme of demographic surveys, commencing with the World Fertility Survey and continued by the Demographic and Health Surveys, has sought to address some of the limitations and restrictions associated with demographic estimation from census or vital registration data. Unfortunately, the Southern African coverage of the WFS was restricted to Lesotho. One official DHS was conducted by Botswana in 1987 (although the data remain strictly controlled and not in the public domain), and an unofficial (i.e. not run under the aegis of MeasureDHS) survey was conducted in 1998 (Central Statistics Office 2001). We were fortunate to be given access to both sets of data, but problems associated with small sample sizes, sampling and data collection mean that these data source are much less reliable than one would

[3] Completeness of vital registration data on deaths in South Africa is estimated to be around 85% (Dorrington, Moultrie and Timaeus 2004; Anderson and Philips 2006)

like: for the most part, the estimates are unreliable and inconsistent with data from other sources in the country. As a result, direct estimation of essential demographic parameters is nigh impossible, and a greater reliance has to be placed on indirect methods of demographic estimation.

With the limitations of not being able to derive most input parameters directly, and with the added difficulties of working with limited and defective data, it is paradoxical that the process of deriving a set of population projections become more important, not less so. Population reconstructions offer a crucially important vehicle for assessing the overall consistency and plausibility of demographic estimations. Without a population projection, there is nothing to stop these demographic parameters from being internally consistent and reasonable, but utterly and evidently incorrect when viewed in the context of their implications for population growth. A case in point from the country under observation: it is manifestly impossible to arrive at the enumerated population aged under 10 in 2001 (undercounted as it was) from the published fertility and child mortality estimates derived from the 1991 census, even allowing for migration.

Methods and data

In order to project the population of Botswana from the aggregation of the 28 district projections, an over-riding methodology had to be decided upon that would be applied to model each district. The complexity of the methodological approach adopted to project sub-national populations in an era of HIV/AIDS from limited and defective data is determined by the rigorousness of the method of allowing for the dynamics of the epidemic in the population(s) being studied; the availability and quality of the data to hand; and the procedure adopted to constrain the sub-national populations to maintain a multi-dimensional consistency with national projections. A wide range of models and methods have been proposed for use in preparing small-area projections (Smith, 2003). The simplest models are those based on trend extrapolation or curve fitting. While they are parsimonious and easy to use, it is impossible to incorporate the detail of the demographic-epidemiological interactions associated with HIV/AIDS epidemics. At the other extreme are multi-state, multi-regional demographic models of the sort proposed by Andrei Rogers (1985). These require extensive matrices of input data, most of which would not be able to be found, or would be spurious, in the settings under consideration, and – again – this approach does not readily permit the incorporation of demographic-epidemiological effects. The requirement to take into account the regional dynamics of the HIV/AIDS epidemic in the country requires that a cohort-component framework be used. Not only are these models the most frequently applied in such circumstances, they also strike the best balance between parsimony and simplicity on the one hand, and complexity requiring a large number of estimated parameters on the other.

Cohort-component models that include HIV/AIDS

Much work has gone into the development of a class of cohort-component projection models that incorporate HIV/AIDS explicitly in their projections. Broadly, there are two approaches. The first produces a projection ignoring the impact of HIV/AIDS and then estimates, usually as a derivation from a possibly independent projection of the prevalence of HIV in the population, the number of people newly infected each year and the number of deaths from the infected survivors each year. These deaths are then removed from the non-HIV projection.

Allowance can also be made, albeit somewhat crudely, for a reduction in the number of births due to HIV and a reduction in the number of AIDS deaths due to the provision of ART. Such is the approach adopted by DemProj and AIM components of the Spectrum suite of programmes developed by the Futures Group (Stover 2005; Stover and Kirmeyer 2005). Their major limitation, of course, is that the estimates of HIV/AIDS incidence, prevalence, morbidity and mortality are effectively decoupled from the demographic projection and there is nothing other than the awareness and vigilance of the user to ensure consistency in assumptions between the two. In addition, since the prevalence is usually the result of curve fitting rather than behavioural modelling the projection can be unreliable and, in addition, it can be difficult to allow for interventions in the fit.

The second approach is to generate the number of new infections as part of the projection model (i.e. firmly embedding the epidemiological model within the demographic model) by making assumptions as to patterns of sexual behaviour and probabilities of transmission. Such models, although more demanding in terms of the number of assumptions that need to be made, are able to allow for a wider range of changes in behaviour and interventions, and do so more realistically.

It stands to reason that the second approach is better, even if significantly more complex, provided one can reasonably estimate the required parameters. One such instance is that employed by the ASSA model (Johnson and Dorrington 2006). This is a multi-state cohort-component projection model, where the states reflect both demographic and epidemiological attributes (e.g. HIV-; asymptomatically HIV+, symptomatically HIV+; on antiretroviral therapy; off antiretroviral therapy; dead from non-AIDS causes; AIDS-related deaths, etc). The number of people becoming infected is determined by splitting the population into four risk groups, each with different heterosexual behavioural[4] and transmission probability assumptions. In addition the model allows for five interventions[5] which make it extremely versatile for modelling the demographic impact of HIV/AIDS, since most countries have implemented some level of these interventions.

The model has previously been applied to project the population dynamics of South Africa as a whole; as well as its nine provinces (Dorrington, Bradshaw, Johnson *et al.* 2006). In particular the model has been used to estimate the numbers orphaned by HIV/AIDS, to cost the provisionof antiretroviral therapy, and more generally a comprehensive national plan to tackle the epidemic, to assess the macro-economic impact of HIV/AIDS, etc.

Ensuring consistency between direct and aggregated models

To ensure consistency between district and national population projections, a combination of a 'top down' and 'bottom up' approach is used to project the population and hence the demographic impact of HIV/AIDS. First, a model which projects the population of the country as a whole is constructed and calibrated to reproduce empirical estimates of the census populations, registered deaths and estimates of the prevalence of HIV as measured by the household survey and various antenatal surveys. This is then used to help construct a prototype model for each of the census districts taking into account district-specific base populations, non-AIDS mortality and fertility rates, and interventions, but not migration. Each district model is then calibrated to fit the district-specific prevalence data. A brief explanation of how the demographic parameters were derived appears in the Appendix.

District-specific migration for each sex, district and age was estimated by first applying the 'ratio method' (a common method of apportioning populations between regions, see for example Shryock and Siegel (1976)) to the projections of the national population from the national model together with the distribution of the population by district, sex and age, at the three censuses. These totals by district and age for each sex separately were then distributed to ages in each district using and expansion of the contingency tables spreadsheet found in the PASEX software package (Arriaga 1994) using the district population numbers by age projected by the prototype district models as the starting point. The resulting district population numbers

[4] Heterosexual behaviour is a function of: the risk groups of partners, the number of new partners per year, the number of contacts per partner, the proportion in which a condom was used, the age of the partners and the sex activity by age.

[5] Information and Education Campaign (IEC), Voluntary Counselling and Testing (VCT), treatment of sexually transmitted infections (STD), PMTCT and antiretroviral therapy (ART).

represent in some sense the population after allowing for migration. An estimate of the number of migrants in 1981 was derived by projecting the district populations forward one year from 1980 without migrants and comparing these numbers with those derived using the ratio method for the population for 1981. These numbers are then included in the model and the population projected to 1982 and compared with the 1982 population derived using the ratio method, and so on until the model includes the number of migrations for each year. The district models with migrants are then recalibrated to the district prevalence data.

The results of these projections are then aggregated to produce a second national projection. Of course, due to the non-linear nature of the models, the results of the two projections cannot be expected to be exactly the same; however, if they differ significantly this indicates a need to improve one or more of the assumptions underlying the models. Once done a national model thus results which corresponds closely to the aggregate of the district models.

There are several reasons for employing such a complicated and time-consuming approach. The first is that modelling the districts allows us to capture, to the extent that the data allow it, the different regional dynamics that there might be in the epidemic, in particular epidemics that started earlier or later that, grew faster or slower than, and/or have reached a plateau above or below that of the national average. The second is that it is more useful from a health (and other) management position to have a model of the local impact of the epidemic. On the other hand, many of the parameters are more reliably estimated at a national level and thus the constraint to have the district projections aggregate to the national projection places a limit on the errors in the district level projections.

By modelling the national dynamics first, one has the additional benefit of working with larger and more reliable data sets: one does not (for example) have to contend with the problem of changing boundary definitions over the projection time frame. An additional benefit is that parameters that are presumed to be relatively invariant between districts (e.g. behaviour of risk groups, various epidemiological assumptions such as survival once infected and the impact of HIV on fertility) can be determined from census or survey data at a national level or elsewhere, and then held constant for each district. This reduces the number of assumed values of parameters required, particularly so when one is attempting to perform cohort-component projections on 28 separate populations.

The conceptual implications of estimating the district-level migration as we have are not immediately obvious, and should be described explicitly. There are in fact two significant problems with this approach. The first is that the changes in the district proportions of the population in the past are determined by not one (as is the case in the absence of an HIV/AIDS epidemic) but two different forces (differential migration and impact of HIV/AIDS epidemic), and by projecting forward using the trend in these proportions one is assuming that the net impact of both remains consistent with the patterns of the past, which may not be the case. The second is that by using the age distributions of the populations subject to district-specific HIV/AIDS epidemics one probably distorts the age pattern of migration away from ages and populations with high levels of HIV/AIDS.

The approach we have adopted privileges the epidemiological aspects over the migratory. Put differently, what is being assumed here is that migration makes up any differences in the populations between districts' differing epidemiological experiences.[6] This is done in preference to assuming a pattern of migration, and then treating the demographic consequences of HIV/AIDS as the balancing item.

Is such an approach is tenable? We think that it is. First, in the absence of any significant medical breakthroughs or interventions (both of which appear unlikely) there is an inevitability (or path-dependence) about the nature and trajectory of an HIV/AIDS epidemic even up to several decades from any given reference point; the same cannot be said about national patterns of migration (let alone that pertaining at a sub-national level). In essence, there is a greater degree of certainty about future epidemiological outcomes than there is about future migration trends, levels and patterns. Second, the demographic impact of a generalised, high-prevalence epidemic (such as that being experienced in Botswana) will usually have a much greater impact on (certainly larger) regional populations than migration is likely to have. In this vein, too, treating migration as the balancing item in the projections, and privileging the epidemiological data makes intuitive sense.

Results

Figure 1 compares the projected population by census district with that produced by the Botswana CSO (2005). Allowing for the difference in the overall projection the numbers are very similar[7] in terms of total population (if not the age distribution), with the exception of Barolong[8].

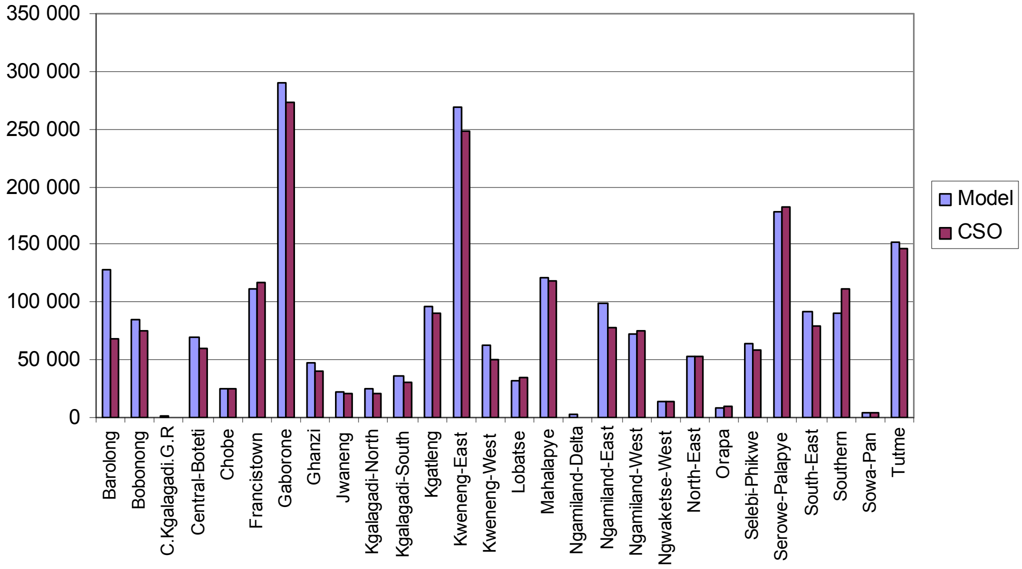
[6] To the extent that the epidemic results in a fall in labour supply and the prospects of employment are a pull factor for migration this assumption is not unreasonable.

[7] In fact in part the close correspondence in the numbers overall despite the fact that the CSO's national projection is some 9% lower than our national projection, seems to be the result of a lack of rebalancing of the district projections to the national projection in the CSO projection.

[8] This is the result of the conscious decision we made not to use extrapolated growth rates for this district as discussed above.

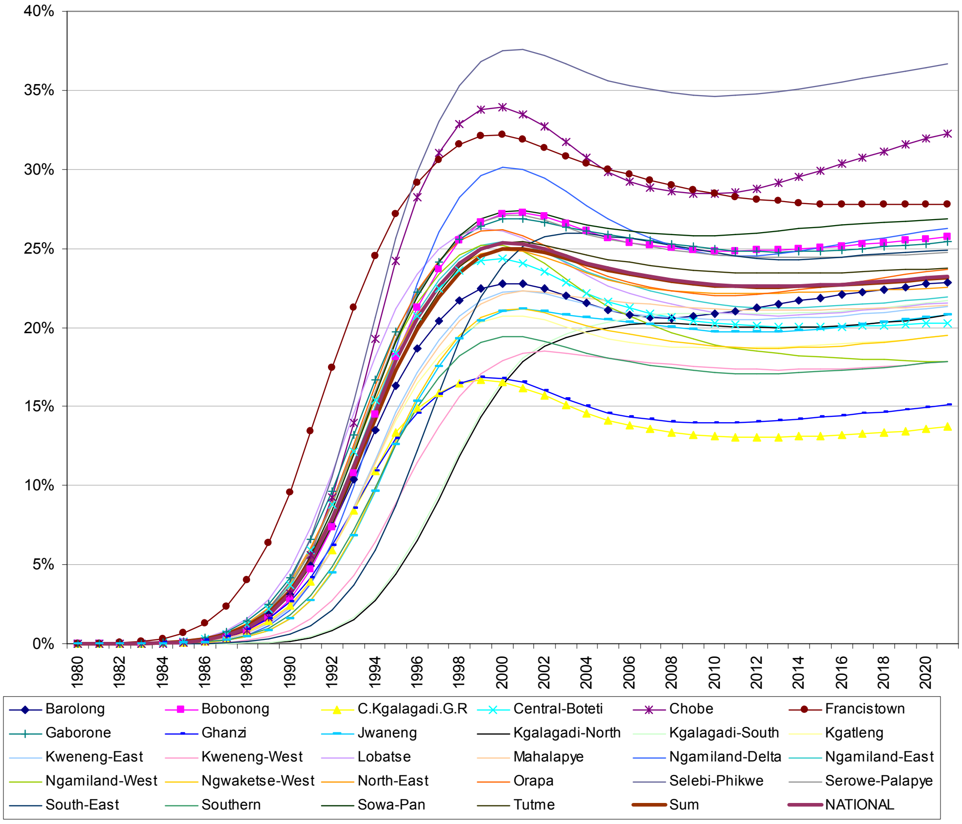
The CSO projection ignores the impact of district-specific epidemics, effectively assuming the same level and timing of epidemic in each district, as one might expect given that migration is the balancing item. Thus we can see that even over a fairly long projection period the method we employed results in little difference in the relative sizes of the overall district populations, but does allow one to assess more accurately the local impact of HIV/AIDS and any interventions going forward.

Figure 1 Population by census sub-district in 2020



As is mentioned earlier, Barolong grew very rapidly between 1991 and 2001 and in projecting forward we reduced the growth rate by 2.5% p.a., which we think was sufficient adjustment, but clearly there is some uncertainty about the future prospects of this district.

Figure 2 Prevalence of adults 15-49 by census sub-district

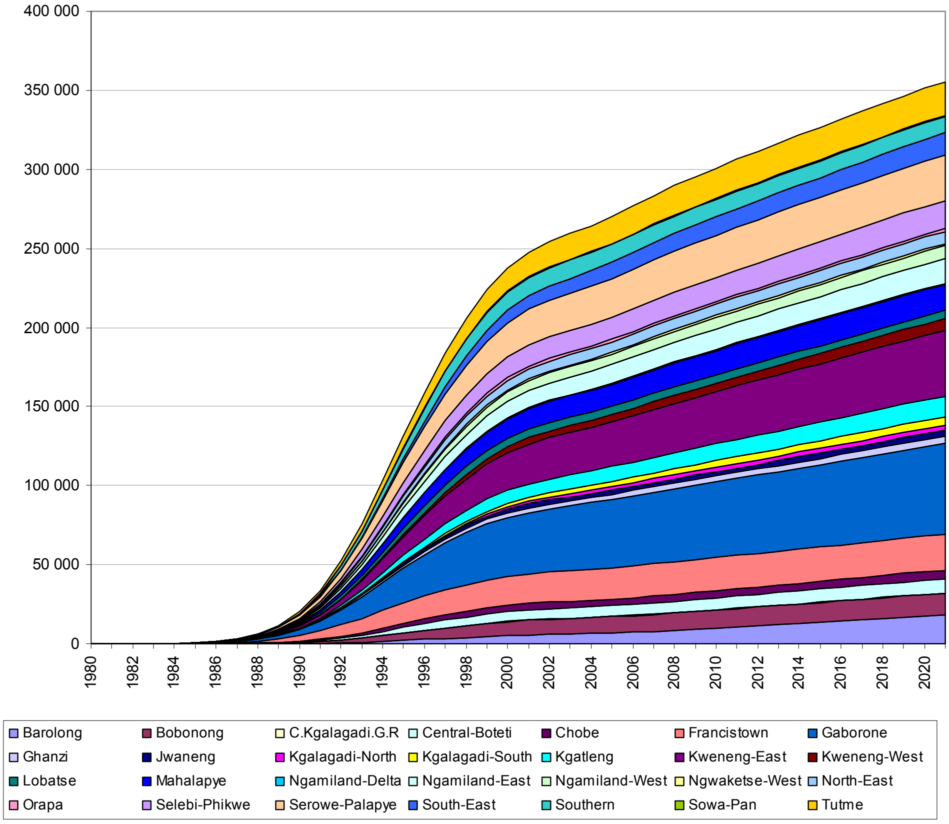


The timing and spread of the epidemic varies considerably between census districts as can be seen from Figure 2, with the epidemic in Francistown being both early and severe (along with Selebi-Phikwe and Chobe) while Kgalagadi North and South are late starters and Kgalagadi Game Reserve and Ganzi are amongst the lowest.

As might be expected (Figure 3) the greatest numbers infected are to be found by and large where there are the greatest concentrations of the populations, Francistown, Gaborone and Kweneng-East. Also, as might be expected, the prevalence rates are expected to remain high and the numbers of infected to increase over time in most districts as more people access antiretroviral therapy.

As indicated earlier these numbers have to be, in aggregate, consistent with the numbers produced by the national model.

Figure 3 Numbers infected by census sub-districts and year



Although the different levels of infection of the various districts result in different impacts on the life expectancy (Figure 4) and under-five mortality (Figure 5) by district, comparison of the estimates in the 1980s shows that some of the difference is due to differing non-AIDS mortality by sub-district, and that, certainly in the case of childhood mortality, the difference is expected to narrow over time.

Figure 4 Life expectancy by census district

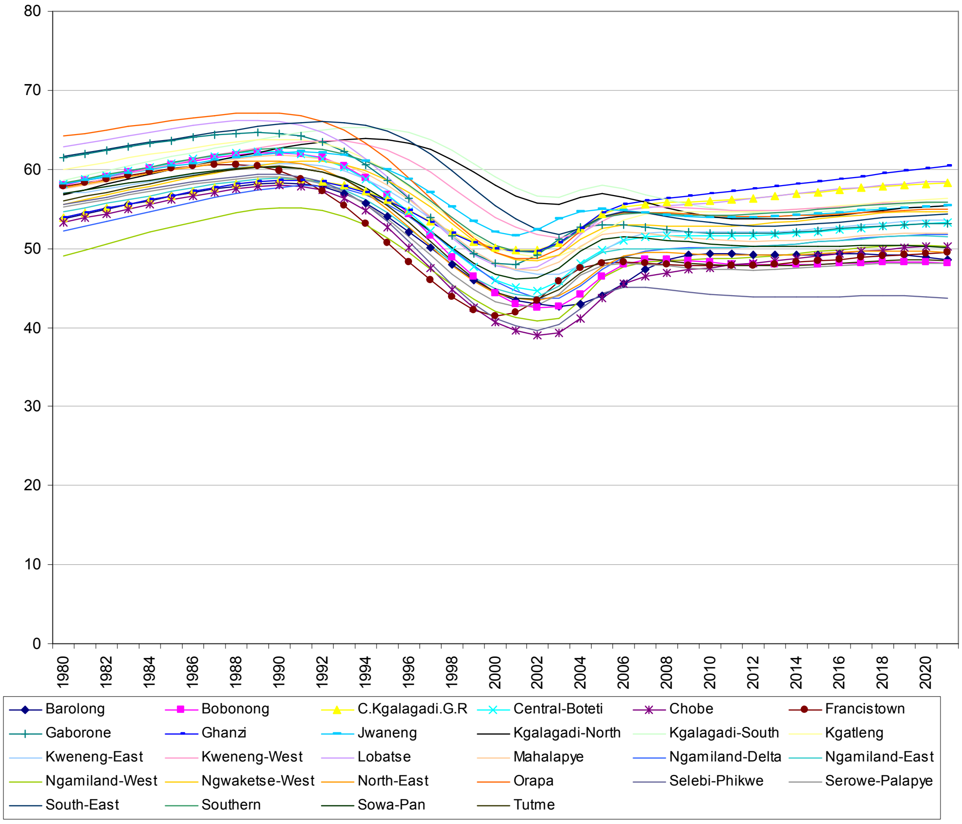


Figure 5 Under-five mortality rates by census district

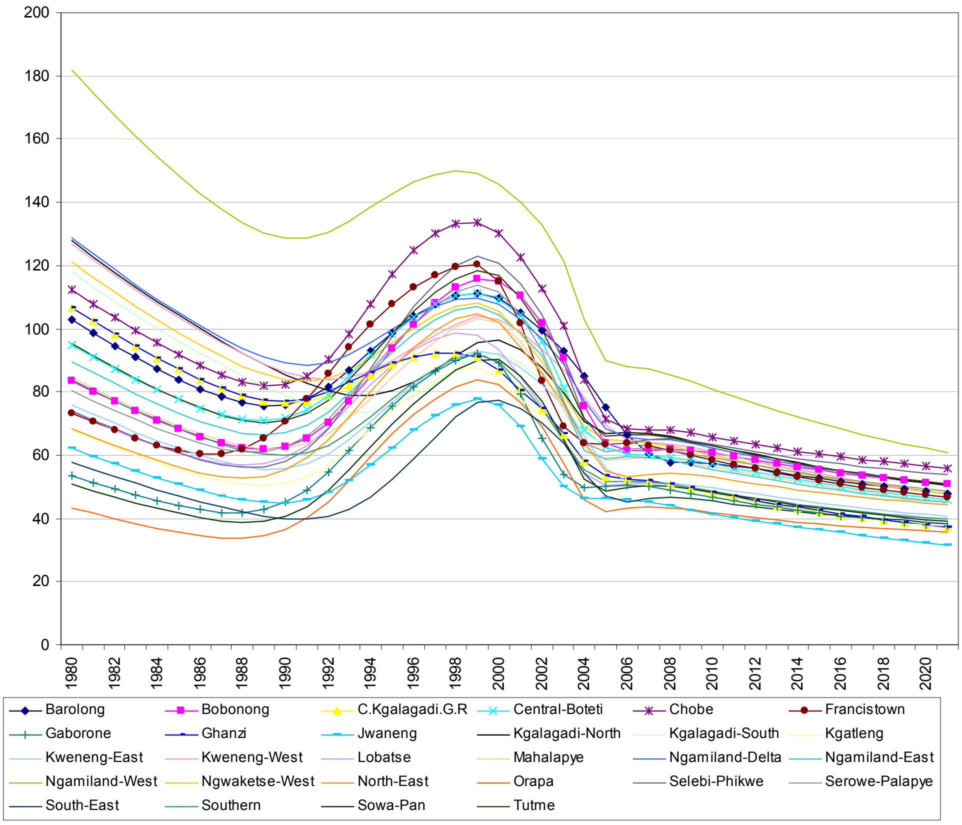
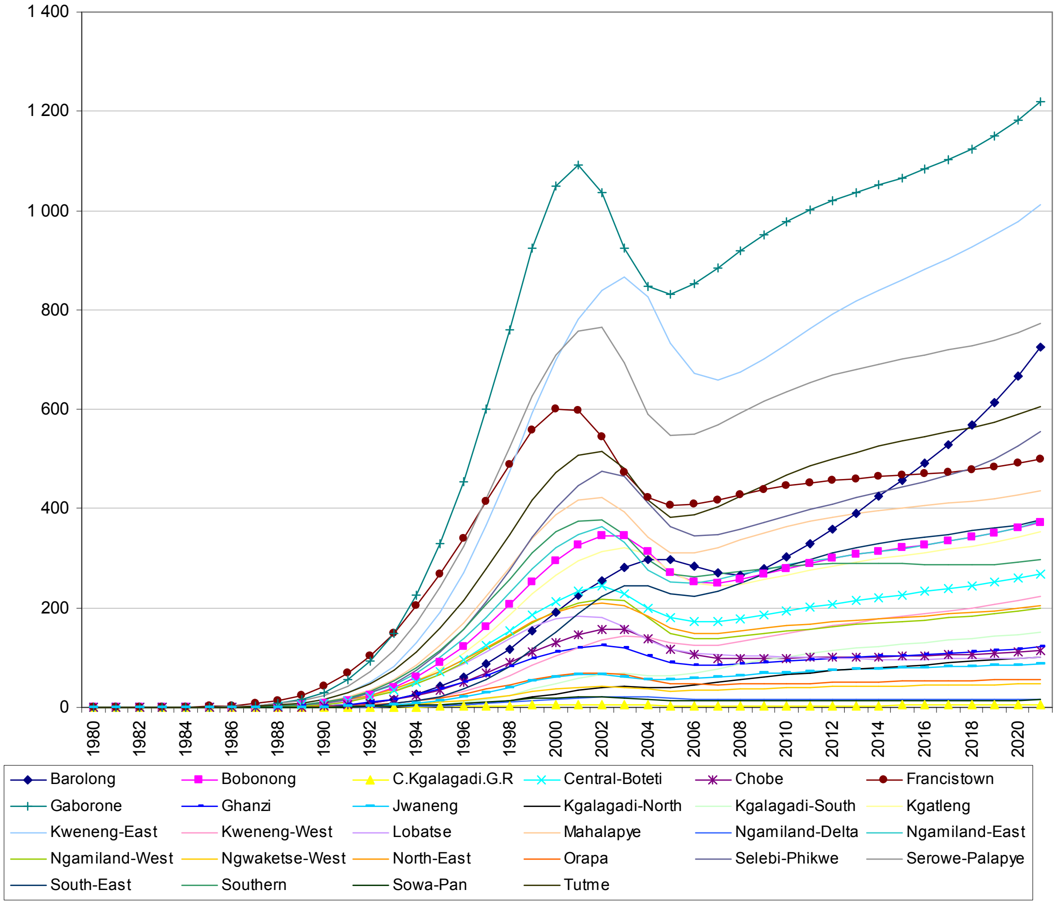


Figure 6 Number of HIV/AIDS deaths by census district



Figures 5 and 6 illustrate that in terms of mortality the districts have been through periods of fairly marked changes against which, if one were to set up appropriate monitoring processes, one would be able to check the reasonableness of the projections and some of the assumptions of the model.

Conclusions

Several important conclusions arise from this exercise in applied sub-national demography. First, that – despite the obvious difficulties and inherent errors involved in a project of this magnitude under such conditions as were experienced – the exercise can provide information, which is not only valuable for policy purposes, but also more nuanced than that which might have been produced by adopting a simpler approach. Second, the use of an iterative top-down-bottom-up approach has been shown to add value in situations where one has to constrain the sum of many sub-national projections to a whole, and has been shown to work when the constraints are more significant than mere aggregations by age and sex.

There is no reason why the approach presented here could not be adapted to any situation requiring constrained projection on any number of dimensions. Third, the approach presented allows the practitioner to establish some parameters iteratively at a national level first, where small-area data may be too sparse or unknown.

An important aspect of the current work is in setting out the reasoning adopted when there are two possible balancing items to be used in completing the demographic estimation, and where (consequently) one has to be privileged above the other: that item which has the greatest long-term impact, and whose impact is more readily known should be preferred. In any sub-regional population projection in Southern or sub-Saharan Africa, it will always and everywhere be the case that more is known about the future trajectory of HIV than about migration.

An alternative approach would be to estimate the migration on the assumption that the level and timing of the epidemic was the same in each district as for the country as a whole and then include these numbers of migrants in the district specific models and aggregate as before. Although there is no reason for supposing that this approach would be any more reliable it does remove the implicit assumption that inter-district migration balances differences between the demographic impact of district epidemics.

In summary, complex population projections incorporating epidemiological dynamics for small areas are possible using cohort-component population projection techniques, even where data are limited and of poor quality, although a case could be made for grouping very small populations to the point where impact of migration becomes less significant than the impact of HIV/AIDS. Such projections can, and should, be used as a tool by decision makers to devise better-informed policies.

Acknowledgements

This research was undertaken as part of a tender called for by UNDP and the National AIDS Co-ordinating Agency of Botswana. Their assistance as well that of the Central Statistical Organisation and the Ministry of Health of Botswana are acknowledged.

References

Anderson, B. A. and Phillips, H., E. 2006. *Adult mortality (age 15-64) based on death notification data in South Africa: 1997-2004*. 03-09-05 (2006). Pretoria: Statistics South Africa. Available: http://www.statssa.gov.za/publications/P03093/P03093.pdf

Arriaga, E. E. 1994. *Population Analysis with Microcomputers. Volume 1: Presentation of Techniques*. US Census Bureau.

Botswana Central Statistical Office. 2005. *Population Projections for Botswana 2001-2031*. Gaborone, Botswana: Central Statistical Organisation. Available:

Bradshaw, D. and Timaeus, I. M. 2006. "Levels and trends in adult mortality," in Jamison, D. T., Feachem, R. G., Makgoba, M. W., Bos, E. R., Baingana, F. K., Hofman, K. J. and Rogo, K. O. (eds). *Disease and mortality in sub-Saharan Africa*. Washington: World Bank,

Central Statistics Office. 1987. *1981 Population and Housing Census: Analytical Report.* Gaborone: Government Printer.

Central Statistics Office. 2001. *Botswana Demographic Survey 1998*. Gaborone: Government Printer.

Central Statistics Office. 2004. *Botswana Statistical Year Book - 2003*. Vol. 28, No. 1.4. Gaborone: Central Statistical Office. Available:

Dorrington, R. E., Bradshaw, D., Johnson, L. and Daniel, T. 2006. *The Demographic Impact of HIV/AIDS in South Africa. National and Provincial Indicators 2006*. Cape Town: Centre for Actuarial Research, South African Medical Research Council, Actuarial Society of South Africa. www.commerce.uct.ac.za/Research_Units/CARE/RESEARCH/PAPERS/ASSA2003Indicators.pdf.

Dorrington, R. E., Moultrie, T. A. and Daniel, T. 2006. *The Demographic Impact of HIV/AIDS in Botswana*. Gaborone: UNDP and NACA, Botswana. Available:

Dorrington, R. E., Moultrie, T. A. and Timæus, I. M. 2004. *Estimation of mortality using the South African 2001 census data*. Monograph 11. Centre for Actuarial Research, University of Cape Town. Available: http://www.commerce.uct.ac.za/care/Monographs/Monographs/Mono11.pdf

Econsult. 2006. *The Economic Impact of HIV/AIDS in Botswana*. Gaborone: National AIDS Coordinating Agency (NACA) and UNDP, Botswana. Available:

Hill, K. H. 1990. "Demographic estimation from deficient or defective data," in Fetter, B. (ed). *Demography from Scanty Evidence: Central Africa in the Colonial Era*. Boulder CO: Lynne Rienner, pp. 133-142.

Johnson, L. and Dorrington, R. 2006. "Modelling the demographic impact of HIV/AIDS in South Africa and the likely impact of interventions", *Demographic Research* **14**:541–574.

Ministry of Health and Child Welfare (Zimbabwe); US Centers of Disease Control and Prevention; UNAIDS. 2004. *Zimbabwe National HIV/AIDS Estimates 2003*. Harare, Zimbabwe: Ministry of Health. Available:

National AIDS Co-ordinating Agency (NACA), Central Statistics Office and Partners, O. D. 2005. *Botswana AIDS Impact Survey II: Statistical Report.* Gaborone: National AIDS Co-ordinating Agency. Available:

Rogers, A. 1985. *Regional Population Projection Models.* Beverly Hills CA: Sage Publications.

Shryock, H. S. and Siegel, J. S. 1976. *The Methods and Materials of Demography (Condensed Edition).* San Diego: Academic Press.

Smith, S. K. 2003. "Small-Area Analysis," in Demeny, P. and McNicoll, G. (eds). *Encyclopaedia of Population*. Vol. 2. New York: Macmillan Reference USA, pp. 898-901.

Stover, J. 2005. *AIM Version 4: A computer Program for Making HIV/AIDS Projections and Examining the Social and Economic Impacts of AIDS*. Washington: The Futures Group.

Stover, J. and Kirmeyer, S. 2005. *DemProj Version 4: A Computer Program for Making Population Projections.* Washington: The Futures Group and the Research Triangle Institute.

UNAIDS. 2006. *2006 Report on the Global HIV/AIDS Epidemic.* Geneva: Joint United Nations Programme on HIV/AIDS/World Health Organization.

United Nations Population Division. 2005. *World Population Prospects: The 2004 Revision CD-Edition - Comprehensive Dataset.* New York: Department of Economic and Social Affairs, United Nations. Available:

Ward, P. and Zaba, B. Unpublished. "The Effect of HIV on the Estimation of Child Mortality Using the Children Surviving / Children Ever Born Technique", Zaba, B. 2003. "Child mortality: development of a functional representation for net mortality due to HIV," Paper presented at UNAIDS Reference Group on Estimates, Modelling and Projections. Sintra, Portugal, 2-4 December.

Appendix: Derivation of the demographic components of the national and district population models

Base population

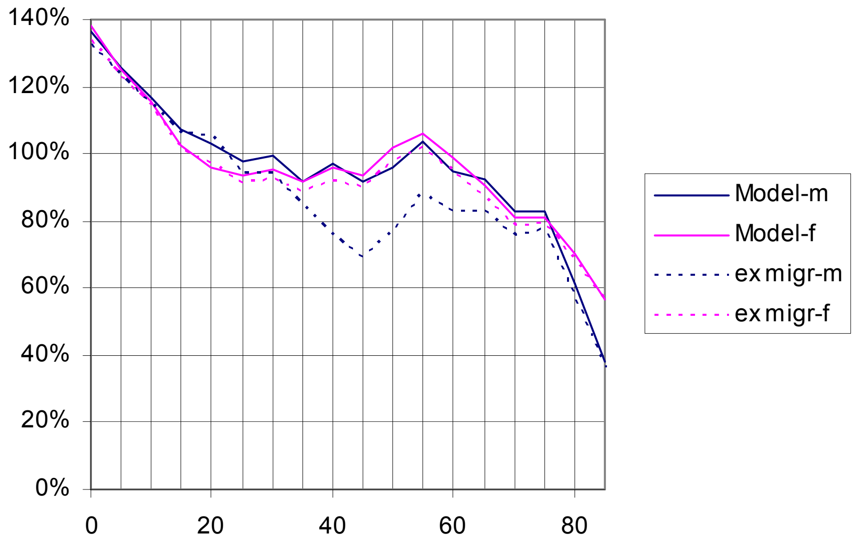
In the case of Botswana, epidemiological evidence suggests that the HIV/AIDS epidemic began in the early 1980s. This requires, then, commencing a cohort-component population projection at least by that point, which in turn requires suitable estimates of the base populations, as well as trajectories of fertility, migration and the single decrement table associated with non-AIDS mortality, both nationally and for each district. 1980 was chosen as the start date to facilitate comparability with other models and the results derived by other researchers (e.g. (UNAIDS 2006; United Nations Population Division 2005)). To set the base population, it was decided to start with the 1981 census back projected one year, which gave rise to the first problem – there are at least four somewhat different tables purporting to be the enumerated population[9]. First, the estimate of the combined total of males and females in each age group was derived by multiplying the ACAP numbers by the ratio of the total population (which was the same for the other three populations) to that for the ACAP data to get a scaling factor. These numbers were then divided into males and females using the five-year ratios for 1981 from the 2001 analytic report. Those of unknown age were also apportioned in the same proportion as those in this report. Since there were clearly far too many people aged "98" it was assumed that this code also represented those of unknown age and these were redistributed in proportion to other ages along with the 99s. These numbers were then back projected one year using the mortality and migration estimates in the model, to give the base, or starting, population.

This base population for 1981 was projected forward (using derived estimates of fertility, mortality, migration and HIV/AIDS) and compared to the census populations for both 1991 and 2001. The base population was then adjusted by a factor determined to be the weighted average of the ratio of the populations of each census at cohort-corresponding ages to that in the 1981 population, using the census numbers in each census as weights. However, for ages in 1980

[9] The four are the 1981 census analytical report 1987 and the 2001 census analytical report both of which are after redistributing those with unknown age to the other ages; the 2003 Statistical Yearbook which gives the number with unknown ages separately; and the 25 per cent sample as archived by the African Census Analysis Project (ACAP) at the University of Pennsylvania, which is the only one available for each individual age, but this is for both sexes combined, and only in total for each sex separately. The last of these sources is the only source giving numbers by individual ages and although the total is not exactly a quarter of the totals for the other sources, and the numbers are particularly different for age groups 75 and above, it was used to derive a distribution of the population by age for each sex in 1981.

of 59 and above for males and 69 and above for females the numbers in the subsequent censuses far exceeded the projected number, and it was decided to assume that this was some form of age exaggeration (puzzling though this is), and hence these numbers were not adjusted further. Figure A.1 shows the comparison of the projected population against the census in 2001. Three things are apparent from this figure. The first is that census appears to have undercounted the number of children under age 20 for males and 15 for females quite significantly. The second is that there appears to be, as was mentioned above, quite extensive age exaggeration. The third is that migration is significant particularly in the case of males. Of course in terms of numbers the difference at the young ages swamps all other differences.

Figure A.1 Ratio of the projected population to the census population in 2001



(Dotted lines represent the ratio ignoring migration; m – male, and f – female)

The base populations and migration numbers for each of the census districts were determined in several stages. The first stage was to use the national model created using the national base population, fertility, mortality, migration, epidemiological and intervention assumptions. Thereafter, estimates of the population by age and sex for each district for each year from 1980 to 2020 were derived as described below.

For each sex, the proportion of males and females in each district was determined from the census. Where the census district population was not reported separately in the census it had to be estimated. In 1991 only Ngwakatse-West was missing and its proportion of the total population was assumed to be the same as in 2001. In 1981 several districts were not in existence or identified separately in the census (Sowa-Pan, Ngwakatse-West, Kweneng-East and West) and these proportions had to be estimated similarly by setting the smallest to the proportions from subsequent censuses and reducing the proportion of the 1981 district in which these populations were assumed to be included. As the proportions are generally quite small, and as process is self-correcting over time, these approximations are of little consequence.

The district proportions of the population for the years between the censuses were derived by exponential interpolation. The proportions for 1980 were derived by extrapolation of the trend in proportions between 1981 and 1991. Proportions beyond the 2001 census were determined by extrapolation of the trend in changing proportions between 1991 and 2001 assuming that the rate of change in the proportions trended to zero over the following 60 years. For each year the proportions were scaled to ensure they summed to 100 per cent. The projected proportions were then inspected for reasonableness and one, Barolong, appeared to be growing unreasonably quickly and so its projected growth rate was reduced by 2.5% per annum.

Application of the district proportions to the projected national population for each sex thus provides the total number of males and females in each district and each year. These were then distributed to five-year age groups by deriving, for each sex and year, estimates of the proportions of the population by age. The proportions by age were derived by interpolating and extrapolating the distributions by five-year age groups in the 1991 and 2001 censuses. Proportions at individual ages were derived from these quinquennial proportions using Beers interpolation, and these proportions were used to derive a starting populations for each district which were then used in contingency tables to produce, for each year and each sex, populations by age such that the sum of the districts by age was equal to the national population in that age, and the sum of all ages in a district equalled the total population for that district.

The fourth stage was to estimate for each district the number of migrants by sex and age for each year. This was achieved by using the district models to project the population one year later assuming no migration for that year and then differencing these numbers from those derived in the second stage to give the survivors. These migrants were then added to the model and the population projected forward to the next year, etc. As a check the sum of the district migrants had to be the same as the national net number of migrants.

Non-AIDS mortality

Child and adult mortality are dealt with separately since the data available, approaches and trends in the rates differ. The results from each investigation are combined to produce a single life table for each of the years required by the projection. Following that, we turn our attention to estimating the rates for each of the census districts.

Infant and childhood mortality

The primary data on infant and child mortality are the answers to the so-called Brass questions, asking mothers to report on how many children they have borne and how many are still alive. Inspection of these data from various sources lead to the conclusion to base our estimates on the census data alone.

Unfortunately the Brass 'children ever borne – children surviving' technique produces results which significantly biased downwards in an AIDS epidemic (Ward and Zaba Unpublished; Zaba 2003) and thus cannot be relied upon to produce estimates of child mortality in Botswana after about the early 1990s.

The projected non-AIDS mortality rates are based on an exponential extrapolation of the trend in the estimates derived from the 1991 census adjusted slightly in the most recent years to remove the expected influence of HIV on the mortality rates. Sex-specific rates were derived by applying the average ratio of the sex-specific rates to the overall rates for each sex for the first three points in time (in order to avoid the impact of HIV/AIDS on this ratio) derived from the 2001 census data[10] to the rates estimated from the fitted exponential curve. Rates at individual ages from 0 to 4 were derived by assuming that the force of mortality followed a hyperbolic curve with respect to age over this age range as determined by these infant and under-five mortality rates.

Non-child mortality

Deriving adult mortality rates was a complex process that involved a number of stages: estimating rates of mortality for the year prior to the census and in the intercensal period by applying the Generalised Growth Balance and Synthetic Extinct Generations methods in combination; comparing these rates against estimates derived by others; removing the impact of HIV mortality in the most recent years to derive a trend in non-HIV mortality; interpolating rates at individual ages from the quinquennial rates; and finally interpolating and extrapolating these rates using exponential trends to the ultimate mortality in ASSA.

Full life tables and projected non-AIDS mortality rates

The rates at all ages were produced by blending the infant and child mortality rates with the adult mortality rates from ages 10 and upwards. Inspection of the 'reduction factors' implied by the fit of the empirical data up to 1996 suggested that if these were used to project the rates into future

---

[10] This was necessary since the 1991 census did not ask about the sex of the child.

years then male mortality would eventually fall below the female mortality. Thus it was decided to project both rates using the average of the male and female reduction factors for 1981-1996.

District non-AIDS mortality rates

Full life tables were produced by first estimating indices of childhood mortality ($_5q_0$) and adult mortality ($_{45}q_{15}$) for each district and then using these and Brass's relational model, with the national life table as standard, to generate full life tables for each district.

Under-five mortality rates were estimated for both sexes combined for each district as reported by women 35+[11] (i.e. in the period 1986 to 1993) in the 2001 census data, using the same method and standard table as used to produce the national rates. These rates were then averaged and compared to the average of the national rates for these women. From this a scaling factor was derived which as assumed to apply to the non-AIDS mortality rates from 1980 to 1996 for both boys and girls.

Since adult mortality is based on the deaths reported by households to have occurred in the previous year or so, the only available data not to be significantly influenced by HIV/AIDS was that from the 1991 census. Thus these data were used to produce estimates of adult mortality ($_{45}q_{15}$) with the intention of then estimating scaling factors from a comparison of the district-specific rates with those of the country as a whole on the assumption that there was no reason to suppose households in one district reported deaths better than households in any other district. Unfortunately, the sample sizes in some districts were too small to produce reliable estimates and rates for these sub-districts had to be approximated by regressing the estimates of male and female rates on the rates of both sexes combined for those sub-districts (the majority) where the results looked reasonable and applying the regression to all sub-districts where the rate for males and females combined looked reasonable. Mortality rates for other districts were approximated from the rates for neighbours in the district. These rates were used to derive scaling factors (the ratio of these rates to those for the country as a whole) and these were used to scale the national non-HIV adult mortality for each district.

Fertility

Estimates of fertility are required, by single year of age, for each year of the projection period from 1980 for each census district. Detailed investigations were conducted to understand the

[11] These older women were chosen to limit the impact of HIV/AIDS on the results.

trajectory of Botswana from 1980 to the present, using as much data from the country as possible, as well as official reports on censuses and surveys conducted over this period.

From this it is apparent that official estimates (as given in the census Analytical Reports for 1991 and 2001) were derived using a methodology that will underestimate fertility rates if children under the age of 5 are differentially undercounted in the census, a feature of the two enumerations conceded elsewhere in those same Analytical Reports. Consequently, the estimates of fertility in 1991 and 2001 (and hence all intervening years too) used in the parameterisation of the fitted model are somewhat higher than the official estimates. Fertility levels are estimated as 5.5 children per woman in 1991 and 4.2 in 2001compared to the official estimates of 4.2 and 3.3 children per woman in the same years.

Individual record data from the 1981 census were not available, so all that was possible was to use published data as inputs into a Relational Gompertz model (unavailable in 1981) to derive an estimate for 1981 of 6.5 children per woman; somewhat lower than the 7.1 suggested in the 1981 census Analytical Report, but roughly equivalent to the 6.6 suggested for 1981 in the 1991 and 2001 reports, although we have found no documentation on the provenance of this figure.

Fertility levels for each district, by individual age and single year from 1980 to 2001, were derived by interpolation from the estimates for each district in 1981, 1991 and 2001. For years after 2001, fertility levels were estimated by assuming an exponential decline in fertility thereafter, taking care to ensure that there were no significant discontinuities in the projected fertility rates. We anticipate fertility achieving replacement levels between 2015 and 2030, depending on the district being considered.

International migration

For most countries international migration is a relatively small component of the national demographic balancing equation. This is not the case in Botswana which has experienced nontrivial flows over the 20 years from 1981 to 2001. In addition to this it is likely that, in common with South Africa, the country will have experienced its share of hidden migration from Zimbabwe, particularly in recent years. Unfortunately, as is the case with most countries, particularly developing countries, it is extremely difficult to document accurately these flows of migrants, and particularly little can be done to estimate the numbers of migrants of which there is nothing other than anecdotal evidence. We adopted a two-stage approach to deriving estimates of the numbers of migrants.

The first stage was to estimate the flow of immigrants net of emigrants by sex and age over each of the intercensal periods by making use of the change in the 'stock' of each of the foreign-born population resident in Botswana, and of the Batswana population resident outside the country between two censuses. The second stage was, as part of the reconciliation of the census populations, to check if there was an excess in the 1991 and 2001 censuses that could be reasonably explained by hidden migration – if so then the numbers of immigrants or emigrants could be adjusted accordingly.

The number of surviving immigrants less the number emigrants in five-year age groups up to the open interval 75+ as at the end of the intercensal period were derived as the difference between the number foreign-born less the number of absentee Batswana at that age in the second census, less the same figure ten years younger in the first census. To these numbers were added back the number that might have been expected to have died before the second census on the assumption that migration took place uniformly over the intercensal period. Numbers of migrants at individual ages were derived from these numbers using Beers formula and then converted to numbers for each of the years.