IHS Information and Insight (Pty) Ltd
Reg. No 1994/007870/07

First Floor, Tugela House
Riverside Office Park
1303 Heuwel Avenue
Centurion, 0157

Phone: +27 (0)12 622 9660
Fax: +27 (0)12 643 1688
www.ihsglobalinsight.co.za
www.ihsglobalinsight.com
info@ihsglobalinsight.co.za

Disclaimer

Estimating Very Low Spatial Populations

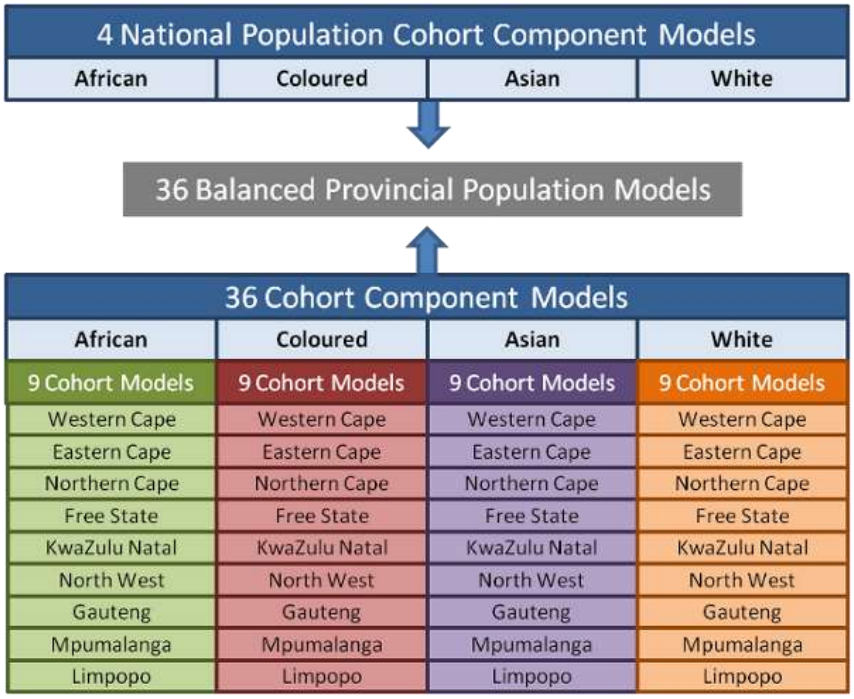
Brief Overview

The population numbers for this project were estimated in four distinct phases. The **first** and **second** phase both make use of the *Cohort-Component Population Projection* method, whereas the **third** phase makes use of the ratio method to break down the results determined by the first two phases into Local Municipal level estimates. Finally, the **fourth** phase distributes the population to the Census 2011 ***Small Area Layers*** (with adjustments for overlapping area type EAs). Cohort-Component Population Projection models are a class of models that are known for their ability to accurately preserve the structure of a population as it grows over time. These models make use of population fundamentals (births, deaths, migration etc.) to project a population as it experiences change.

Phase one: We estimate 4 *national* cohort component models, one for each population group in South Africa. Each group has its own set of assumptions regarding fertility, mortality, migration and so on.

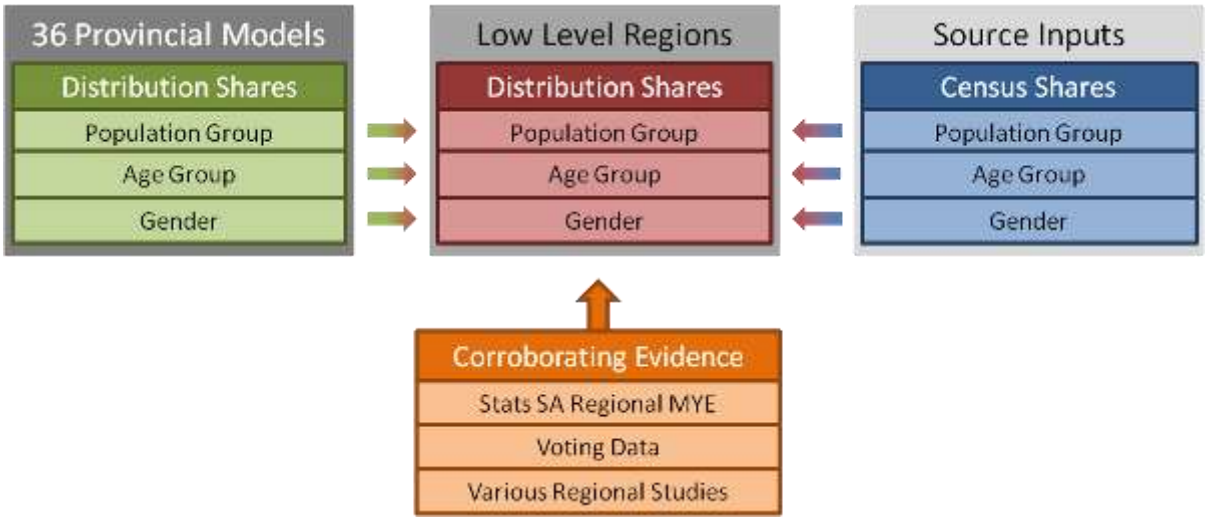
Phase two: We estimate 4 *provincial* cohort component models for each population group in each province. This results in 36 additional models. These 36 models are run numerous times, with subtle changes in their underlying inputs, until the sum of all 36 provincial models is equal to the sum of the four national models (from phase one.)

Phase one and two are depicted in the following image. The grey block (*36 Balanced Provincial Population Models*) represents the output of these two phases.



Phase three: Using distributions from the Census years, we distribute the population to different regions, using a share-of-province, population-group, gender and age-category breakdown. This forces all regions to balance up to the 36 Balanced Provincial Models for each population group, gender and age category. For the years between Censuses we use interpolation, with occasional adjustments where the evidence supports it.

Phase three is depicted in the following image. The grey block (*Low Level Regions*) represents the final output. The block on the left (*36 Provincial Models*) is simply the output from phases one and two. The *Source Inputs* block on the right is determined from the Censuses.



Phase 4 This part of the project is not available 'off the shelf', so the methodology needed to be customised to meet the requirements of the project. We used a multi-model approach to this problem – in an attempt to derive very similar numbers using different techniques – and then weighted each model to account for its strengths and weaknesses. Throughout this phase, we maintain the output from phase 3 as a set of control totals.

Finally, it should be borne in mind that all of this work remains consistent with the broader IHS Regional eXplorer socio-economic model. This data rich model helps to constrain and sanity check the final population outputs on a local municipal level across a much broader spectrum of indicators (economic activity, income, human development etc.)

0. Introduction to Cohort Component Models

The national and provincial population estimates are obtained using a *Cohort-Component Population Projection*. These projections are determined by five fundamental population variables:

- Size of population in the base year, P^t
- Number of deaths occurring between the base and projected years, D^t
- Number of births occurring between the base and projected years, B^t
- Immigrants arriving in the country between the base and projected years, I^t
- Emigrants leaving the country between the base and projected years, E^t

The above variables contribute to the projected population, P^{t+1}, within the constraints of the following demographic balancing identity:

$$P^{t+1} = P^t + B^t - D^t + I^t - E^t$$

The final population figures are based on a set of cohort-component models, with a separate balancing equation for each population group and province. We sum the individual results to arrive at the total national population. This is done because fertility, mortality and migration factors vary largely between South Africa's major population groups, *and* across South Africa's provinces. This methodology ensures an accurate representation of the grouping breakdowns within and across the country and is ideal for projecting populations beyond the data edge.

The total national population itself is derived from an additional four cohort-component models, one for each population group. This ensures that population projections on a national level are realistically captured.

We used Spectrum to run the population projections. The inputs were obtained using various models and assumptions, as discussed below. Due to the nature of correctly balancing and benchmarking a population model, many of these steps are carried out simultaneously and not necessarily in the following order.

1. Determine the Base Population (P^0)

The National Base Population

As can be seen from the demographic balancing equation, the accuracy of P^0 determines the accuracy of population projections made for all P^{0+x}. Vital to the projection however, is the age structure at time t = 0, which plays an important role in determining all age structures for P^{0+x}. The age structure has *significant* knock-on effects with regards other age-dependent drivers such as fertility and mortality.

We estimate P^0 based on results from Census 1970, with additional benchmarking using later Censuses for backward extrapolation. Census 1970 was chosen as a starting point due to the general consensus on this work being of a high standard and level of accuracy. Furthermore, this Census included the previously named 'National States' of Ciskei, KwaZulu, Gazankulu, Lebowa, Qwaqwa, Kangwane, Kwandebele, Transkei and Bophuthatswana - in other words, all of South Africa – which Censuses taken after 1970 and before 1996 did not.

Using the 1970 Census results as a base, and all later Censuses as bench markers (with a weight given to each according to its quality) factor based backward extrapolation was used to arrive at an adjusted 1970 base population. In other words, we ran our cohort-component model with subtle changes in input assumptions until the base population was close to the Census 1970 results and the total population over time was close (depending on the quality of the census) to other Census releases since then.

Our various input assumption adjustments, and our adjustments to the base population, were made within a range of expected norms until we arrived at the final output for the 1970 national population. This process was carried out for each population group individually, resulting in four unique base populations.

It should be noted that most Census results are adjusted *ex-post* by various academic researchers in order to maintain time consistency with existing expectations and previous Censuses. In determining the base population, we used the adjusted Census 1996 and adjusted Census 2001 estimates which were built on work done by Prof. Rob Dorrington of the Actuarial Society of South Africa and the University of Cape Town.

The Base Population for Each Province

Once the national base population had been estimated, the same exercise was carried out for each province (and each population group in each province) to arrive at the provincial base population estimates. Naturally, these were calibrated with the constraint that they should sum to the national base population estimate, with special care being taken to ensure that each provincial age and gender distribution matched theoretical and empirical norms.

Although the Census 1970 estimates are implicit as base years in our provincial projection, our provincial models are not explicitly modelled on the 1970 base year. The provincial Cohort-Component models begin in 1996 and therefore use the adjusted Census 1996 estimates as their base year population. This decision is made due to province level boundary changes that are impossible to untangle in censuses taken prior to 1996.

2. Determine the Various Fertility Rates

Cohort-component population models require two different fertility inputs; the first is the *Age Specific Fertility Rate*, defined as follows:

$$ASF_x = \frac{\text{Births in year t to women aged x last birthday at time of birth}}{\text{mid year population of women aged x last birthday}}$$

The second fertility input is the Total Fertility Rate, which is merely a sum of the Age Specific Fertility rates across a single average woman's entire span of possible birth years (in other words, the fertility rate between the ages of 15 and 50 added together.)

Total Fertility Rates therefore represent the average number of children that a woman in the target population will bear between the ages of 15 and 50, and is defined as follows:

$$TFR = \sum_{x=15}^{49} ASF_x$$

One way to think about the difference between these two rates is to think of the TFR as the total number of children an average woman will have during her entire life, and the ASFR as the ages at which she will have those children.

Considering the above, we need to estimate TFRs and ASFRs for each population group in the national model, and then each population group in each of the provincial models.

National Fertility Rates

Total Fertility Rates were determined per population group to account for the large differences between each group. These rates are estimated in three stages:

1. Literature review of published demographers' work. These provide starting points for our demographic models.
2. Additional calculations based on Censuses (and other surveys where available.)
3. Calibrating and benchmarking the model against empirical outputs.

In reviewing the literature, it was found that there is broad consensus regarding fertility rates for the **Asian** population.

- Sadie (1993) projected fertility rates for this population to be at 1.8 for the period 2006 to 2011.
- Calitz (1996) projected a TFR of 1.81 for 2015 to 2020.
- van Aardt (BMR publication 272) projected a TFR of 1.7 in 2006, moving down to 1.45 in 2016.
- The standard ASSA 2008 output TFR was slightly lower than the rest, at 1.45 for 2006, moving down to 1.37 in 2016.

There is less consensus regarding fertility rates for the **African** population. Nonetheless, demographers tend to agree on two ideas regarding this population group. Firstly, that fertility rates are decreasing and secondly, that this population has higher fertility rates than the other population groups. These findings are supported by Sadie (1993), Caldwell and Caldwell (1993), ASSA 2008 and van Aardt (BMR publication 272.)

The factors driving the decline in TF rates for the African population group have often been explained by increased urbanisation, a greater use of contraceptives, lower fertility preferences and growing labour force participation rates among women. Many demographers expect that greater urbanisation will lead to even lower fertility rates in this population, as supported by SADHS (1999) which showed that fertility rates among the *urban* African population were 40% lower than the average.

The **Coloured** population group is expected to experience a similar decrease in fertility rates, although less severe, for similar reasons. This is supported by the same various publications.

Fortunately, there is also some consensus in the literature regarding fertility rates for the **White** population group. Specifically that it will remain stable over the period 1996 up to 2020.

- Sadie (1993) indicates that the White population group will experience a decline of TFR to about 1.66 by 2001. This was confirmed almost exactly using calculations from Census 2001.
- Calitz (1996) indicated a progression from 1.7 to 1.5 from 2000 to 2020.
- van Aardt (BMR publication 272) indicates an almost identical progression, although ending slightly lower (1.43) in 2020.
- van Aardt and van Tonder (1999) also support a stable fertility rate in the White population.

Estimates of fertility rates prior to 1985 made use of Udjo (2003), StatsSA's Mid-Year Estimates (2004) and van Aardt (BMR publication 272) which were in broad consensus. The one notably different view, coming from Udjo (2003), was regarding the African population, wherein it was argued that fertility rates had been overestimated in the past. We tended to agree with this view, although to a lesser extent.

Using the above fertility rates as a starting point, and combining those with additional calculations from various census results, we derived a set of final input TFRs. Some trends were derived from various StatsSA household surveys and Mid-Year Estimate assumptions. The final national TFRs were estimated during the calibration phase of the model such that the population estimate started at the given population in 1970 and passed through each of the population figures from the 1985 census up until the latest census, within an adjustment factor that recognised the quality of each individual survey.

The national input Total Fertility Rates are captured in the following table:

Year	African	White	Coloured	Asian
1996	3.29	1.55	2.90	2.05
2001	2.95	1.52	2.60	1.76
2006	2.82	1.63	2.54	1.78
2011	2.84	1.82	2.63	1.90
2016	2.71	1.89	2.57	1.89

Age Specific Fertility Rates were based largely on van Aardt (BMR publication 272.) We carried out some additional smoothing and estimated our own ASF rates for the Coloured population group.

We also considered other publications, including the Standard UN Sub-Sahara profile ASF rates as a comparison. However, these were not used in the final inputs.

Provincial Fertility Rates

Provincial fertility rates are also calculated in three stages:

1. Calculate absolute provincial fertility rates from Census datasets.
2. Calculate the ratio between the provincial Census rates and the national rate for Census years.
3. Calibrate the provincial rates by holding all other inputs stable and benchmarking the sum of the provincial output population to the national population.

   As a constraint, this last step makes use of the StatsSA Mid-Year Estimate provincial fertility rates, and the ASSA 2008 standard provincial fertility rates, ensuring that the outputs from our calibration do not differ significantly.

The output of this methodology results in 36 unique TFRs, one for each population group and province. This is done because, even within the same population group, some provinces exhibit higher fertility rates than other provinces. As an example, one could consider births in the African population group in the Eastern Cape, where there are more children per adult compared to the same population group living in the Western Cape.

This variable was benchmarked to fit the total number of babies born for the period 1991 to 2011. It should also be noted that the 0-4 age category is often underreported during population censuses, and therefore this figure was adjusted slightly upwards to correspond with the national TFRs. Finally, provincial TFRs were adjusted for accuracy in order to calibrate the provincial models to the total national model for each population group separately. This is seen as a more robust method in that it maintains the population group-specific age structures.

3. Determine the Various Birth Ratios

Birth (or sex) ratios measure the number of males in the population per the number of females in the population. Population growth depends largely on the number of females, and this input will therefore determine the overall growth rate of the population. It is possible to measure a Gross Reproduction Rate, which is similar to the TFR, but only measures the number of daughters a woman is likely to have. This is calculated similarly to the TFR, as follows:

$$GRR = \sum_{x=15}^{49} F_x{}^{d}$$

Note that **d** denotes only daughters born to a female over her child bearing years, 15 to 49 inclusive.

As with TFRs, the GRRs differ from one population group to another – and even from one province to another, albeit in a much narrower range. These rates are typically dependant largely upon genetics and are therefore very stable over time for a given population group and province.

The various birth ratios were calculated from a number of empirical sources, specifically the StatsSA Censuses, but also various other population and household surveys. We calculated the input birth ratios as the number males per 100 females. The following birth ratios per population group are used on a national level for 2010, bearing in mind that these remain stable over a long period of time, although we do allow for small changes over time where supported by the data.

African	White	Coloured	Asian
99	102	105	106

In order to determine birth ratios per province, a methodology similar to that used for Total Fertility Ratios was applied. In other words:

1. Calculate absolute provincial birth rates from Census datasets.
2. Calculate the ratio between the provincial Census rates and the national rate for Census years.
3. Calibrate the provincial rates by holding all other inputs stable and benchmarking the sum of the provincial output population to the national population.

It was assumed that differences in birth rates across provinces remained fairly stable over time. Note that this does not imply that birth ratios across provinces remain stable, only the percentage difference (for the province) from the national total. This was again benchmarked against the national level data using the StatsSA Census and the national population estimate.

4. Determine the Various Life Expectancies

Determining average life expectancy is complicated by a number of factors. Life expectancy varies across different genders, population groups, age groups and geographic regions. Furthermore, the

effect of HIV and AIDS on the mortality rates across the population groups is likely to complicate the estimation. This final complication is the point of departure between demographers.

There are two broad routes that demographers can take when estimating life expectancies in a population affected by HIV and Aids. The first option is to estimate life expectancies including the effect of HIV and Aids. In other words, the demographer will estimate life expectancies that are lower than they would have been if there were no HIV / Aids effects in the population. The second option is to estimate life expectancies ***without*** the effect of HIV and Aids, and account for the problem *ex-post*.

Regardless of the approach, the methodology employed should aim to correctly maintain the age-distribution of mortality rates, because getting these wrong will not only affect the structure of the population, but also the overall population size. Therefore, should a demographer account for HIV / Aids by simply lowering the input life expectancies, a model life table that *includes* the effect of Aids must be used. Demographers who account for HIV / Aids *ex-post* can make use of the standard model life tables - but they must build an accurate model of the Aids virus, how it spreads, and how it affects *output* life expectancies.

In this demographic model, we have opted to take the second route, accounting for the effect of HIV and Aids *ex-post*. This decision is based primarily on the fact that the only model life tables that have attempted to incorporate the effect of HIV and Aids are the INDEPTH model life tables and, unfortunately, these tables suffer from the following problematic characteristics:

- Mortality data is collected from various countries across Africa and combined into a single set of Model Life Tables. The implication here is that the various strains of HIV are aggregated into a single mortality distribution, contrary to medical evidence showing that different HIV strains have different progression rates. South African cases of HIV are almost all of the **C** strain, which progresses slightly slower than those strains found in East Africa.
- Model Life Tables are designed to capture a single, static description of mortality in a population. However, contagious viruses (such as HIV / Aids) are not a static event. Different countries, and even regions within countries, are at different stages of the virus. Furthermore, the dynamic state of ARV uptake is continually changing the age distribution of mortality. This is an effect that cannot be captured in a static model life table.
- The current implementation of the INDEPTH life tables is still fairly new and therefore contains significantly less data points than the existing and established model life tables upon which the non-HIV / Aids counterparts have been built.

This model therefore makes use of input life expectancies excluding HIV / Aids, and accounts for the disease *ex-post*. The specific procedure for arriving at the input life expectancy estimates is carried out in various steps, as follows:

1. Literature review of input and output life expectancies per population and gender group.
2. Individual calculations of input life expectancies per population and gender group.
3. Calibration using various constraints:
    - Line of best fit through expected *output* life expectancies. These are obtained after running the model and adjusting the input life expectancies for Aids.
    - Line of best fit through Census population estimates.
    - Total output deaths calibrated to the national deaths register data, after adjusting for underreporting.

The above procedure is carried out for each of the population groups separately, as pointed out above, and thus results in four individual cohort-component models on the national level. These life expectancies are inherited by the appropriate population group in each province, although they are implicitly adjusted when calibrating the 36 provincial models to the national population output.

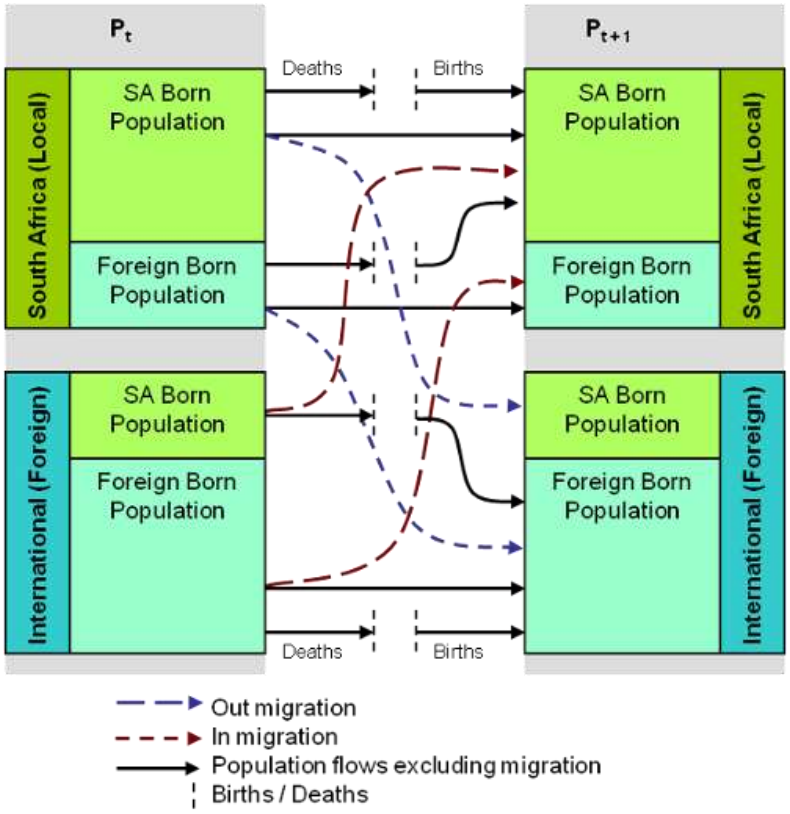
One final note on estimating life expectancies is the choice of *Model Life Table*. Given that we apply input life expectancies that exclude the impact of HIV \ Aids, this model is free to use the standard Coale-Demeny model life tables, or even those compiled by the UN. This is standard practice for most demographers (although it should be noted, not *all* demographers follow this approach.) We use a custom combination of various Model Life Tables dependent upon the population group and occasionally mixed depending on gender within the group.

5. Estimate National Migration Flows

In order to accurately capture the international migration inputs to our cohort component models, we developed a specific South African migration model which was used to compute the inputs before passing them onto the main model. This international migration model was defined by starting with the basic population balancing equation;

$$P^{t+1} = P^t + B^t - D^t + I^t - E^t$$

For the purpose of better understanding migration, the standard balancing equation was expanded to explicitly include all migration flows. The following illustration details the possible migration flows in a population.



The above image depicts a growing population as it moves from time **t** to time **t + 1**, depicted by the large grey blocks on the left and right of the image. This is similar to the standard population balancing equation, but with more detail. Specifically, we make a clear distinction between foreign-born residents and local-born residents. We also consider the foreign population (represented by the bottom blocks on the left and right.)

We illustrate the fact that there are South African and a foreign born individuals residing both inside and outside the country. These individuals are free to move as they please, and to come and go from their birth and resident countries at will. The following migration flows are therefore possible:

- Foreign born population moving into the country.
- Foreign born population moving out of the country (either returning home or moving elsewhere.)
- Local born population moving out of the country.
- Local born population returning to the country after being away.

The balancing equation is thus expanded as follows:

$$P^{t+1} = P^t + B^t - D^t + (I^{Ft} + I^{Lt}) - (E^{Ft} + E^{Lt})$$

Where

- The standard variables retain their meaning.
- I^{Ft} = In migration of foreign born population.
- I^{Lt} = In migration of local born population.
- E^{Ft} = Out migration of foreign born population.
- E^{Lt} = Out migration of local born population.

Furthermore, and also demonstrated using the above image, the following equation is defined:

$$P^{Ft+1} = P^{Ft} - D^{Ft} + I^{Ft} - E^{Ft}$$

From here, we are able to derive a net migration figures, which was done in two phases. We first estimated the change in foreign-born population by considering internal South African surveys, and then the change in the local-born population by considering external (to South Africa) data sources. This is discussed in more detail below.

Where

- The standard variables retain their meaning. P^{Ft} = Foreign born population living in South Africa

Net Migration of the Foreign Born Population

By measuring the change in P^{Ft} we were able to derive a net migration figure for the foreign born population of South Africa. This was achieved by measuring the change in size of the foreign born population between all available censuses – given that we know censuses between 1970 and 1996 did not accurately measure the entire country.

These figures were also compared to refugee statistics from the UNHCR, which are based on asylum applications. The UN also measures the foreign born population living in South Africa, and this measure was considered as well. Other, anecdotal sources of information (such as number of refugees housed by South Africa during xenophobic attacks and the number of deportees monthly by the Department of Home Affairs) were also considered.

Additional evidence was obtained by measuring out migration from the foreign perspective where it was available.

Net Migration of the Local Born Population

Net migration of the local population ($I^{Lt} - E^{Lt}$) was measured largely from the foreign perspective. Data on outward travel suggests that most South African emigrants move to the following five

countries; England, Australia, New Zealand, America and Canada. This is confirmed by the United Nations Migration tables for South African born individuals living in other countries.

We therefore measured immigration of South Africans to the major five destinations, from the view of the migration destinations. In other words, we focussed on work produced by those five destination countries, which fortunately all maintain very good statistical records. We balanced among the following variables for each destination country:

- Change in the South African born population between censuses.
- Long term work permits issued to South African born people each year.
- Citizenships issued (applying the correct number of years of required residence per country to arrive at an accurate migration year.)
- Number of arrivals declaring their intention to migrate.

The accuracy of these findings was confirmed by the proximity of each measure to all of the others. Even when using different methods to calculate outward migration for South Africans, the results were very much the same for the top five destinations. A final adjustment was made for the remaining countries of the world, with adjustment factors confirmed by other foreign perspective data where available.

Age, gender and population group breakdowns are important for a good cohort component model. Fortunately, many of the destination countries did capture such information, and we were therefore able to use the foreign perspective approach to accurately reflect the population structure of emigrants as well.

Recall that four national models are built, one for each population group. We therefore complete the above process for each of those groups, resulting in a total of four net migration flows - all of which make up the four national models.

6. Estimate Intra-National Migration Flows

The national model requires four net migration streams, one for each population group in the cohort-component model. Likewise, each of the nine provincial models requires four net migration streams. It is impossible to ignore provincial migration streams in South Africa. This is due to the structurally changing population across the country. Specifically, since 1994, most evidence (Censuses, voters' roll data, satellite imagery etc.) show that cross-province migration has increased rapidly.

We use the various Censuses and large scale household surveys to measure inter-provincial migration. All migration streams are broken into population groups, gender and age cohort, with a starting input defined using a linear trend between Censuses. Additional assumptions are made for specific cohort groups where the results are unlikely or impossible.

We take the Census 1996 dataset to represent a structural break for migration, being the first census after the regime change and are careful not to use this in our trend analysis of migration. Additional assumptions are therefore required to estimate cross-province migration. We take the following datasets into account when considering our final cross-province migrations figures:

- Change in language share per province over time (because people don't typically change their home language very often, even if they move to a new province.
- Detailed voters' roll data from the IEC.
- Economic factors (which have been shown to be related to the perception of a potential immigration region.)
- Satellite photography showing the growth in dwelling units over time.

We also distinguish between international migration into a province and local migration into a province. This is done using the share of foreign-born residents per province, obtained from the Census datasets and balanced to our national immigration estimates.

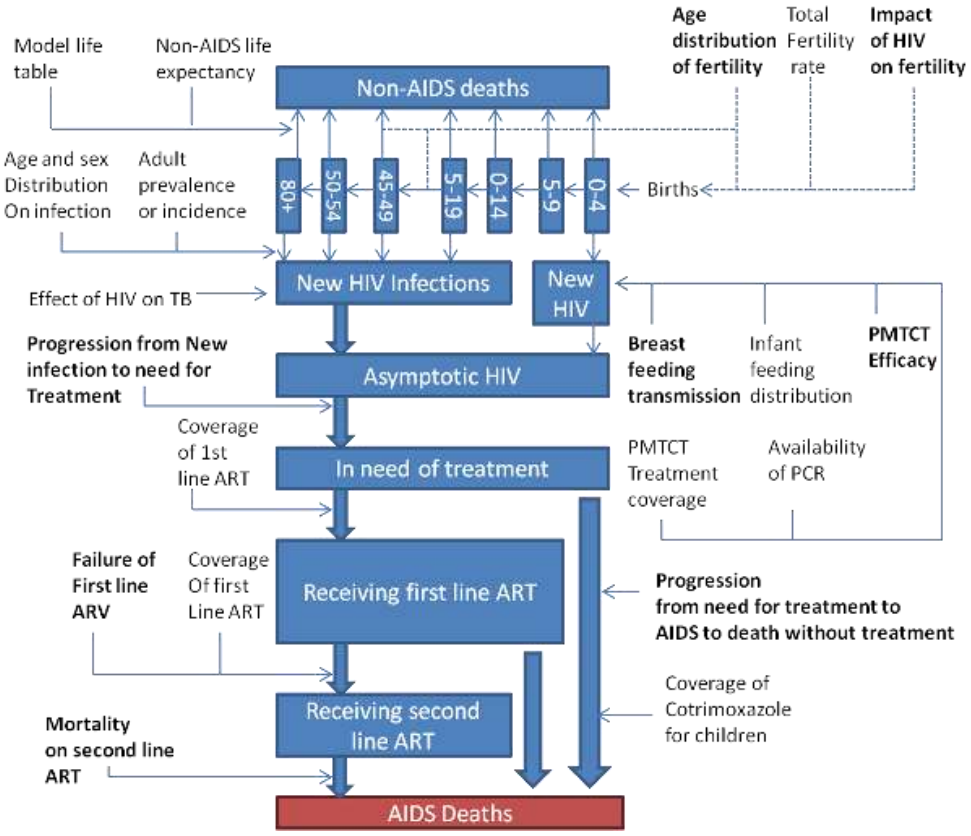
There is some evidence of conglomeration of populations in South Africa. At the same time, there seems to be a relative de-population of the rural areas and small town areas. As an example, whilst the Pretoria-Johannesburg area is experiencing a population explosion, other areas (like Welkom) are experiencing population declines.

7. Adjust for Aids - The HIV / Aids Model

As discussed in a previous section '4. Determine the Various Life Expectancies', we make use of what demographers call *input* life expectancies that ***exclude*** the effect of HIV and Aids. The implication of this methodology is that we can use the standard, high quality model life tables. The downside, however, is that we need to build a separate model to account for the impact of HIV and Aids. For this purpose, we make use of the Aids Impact Model, which is known simply as AIM.

AIM was first developed by *The Futures Group* and *Family Health International* (1991) and has been revised a number of times since its creation. This was undertaken with the help of the UNAIDS Reference Group on Estimates, Models and Projections. The model is currently maintained

and updated with support from USAID. The following image depicts the various inputs and workings of the model:



The AIM approach to Aids adjustments is useful for the following reasons:

- AIM takes, as its main input, an existing cohort-component demographic projection (prepared with non-Aids life expectancies) and adjusts the output accordingly.
- AIM allows the model builder to specify a wide range of assumptions, giving one the ability to very specifically account for Aids impacts in various different and unique populations.
- AIM has proven to be a reliable approximation of reality in a broad range of circumstances.

AIM accepts a number of input assumptions; the following sections discuss briefly our methodology for determining selected, important, input assumptions.

Adult HIV prevalence rates

These rates were obtained from the HIV/AIDS model built by the Actuarial Society of Southern Africa (ASSA, 2008). These rates were input as base starting points. These prevalence rates were adjusted slightly to account for the Model Life Tables we used. We also adjusted the provincial

prevalence rates slightly to reflect the national HIV prevalence output per population group as used in our national demographic models.

The ASSA prevalence rates are based on significant work by ASSA on various primary data sets - particularly the Ante-Natal prevalence surveys conducted by the Department of Health. We also consider the HSRC household surveys on HIV / Aids.

Speed of Progression

The speed of progression (how quickly the HI virus becomes AIDS) is well documented and fairly stable from one population to the next. This does change under differing assumption of Anti-Retro Viral (ARV) rollout, which we need to estimate separately. For the base progression rates, we used a combination of the ASSA 2003 progression assumptions and the UNAIDS slow progression rates for South Africa. We made slight adjustments, particularly to the progression speed after 19 years of HIV. We increased the progression, compared to ASSA 2003 and UNAIDS.

Prevention of Mother to Child Transmission

A number of assumptions on the spread of the disease from mother to child were required. Specifically, we differed from the standard UN estimates for South Africa on the following inputs:

- Percentage of mothers receiving PMTCT treatment
- Percentage of children with moderate to severe HIV disease receiving ART
- Percentage of children born to HIV+ mothers receiving Cotrimoxazole
- Adults with advanced AIDS symptoms receiving ART treatment

We based our estimates for these inputs largely on Dorrington (2006).

Age and Gender Distribution

These estimates were taken from the StatsSA 2004 Mid-Year Estimates, with no adjustments. These figures were compared to the UN estimates and were generally comparable. The StatsSA figures have a slightly higher risk for ages 20 – 24, which fits the South African assumption a little better.

8. Ratio Method for Sub-Provincial Regions

There is not enough information to build cohort-component demographic models on a sub provincial level. Particularly, migration estimates become difficult to derive, and therefore, the

level of accuracy required for this model is not attainable. The alternative technique, if one has both top level estimates and detailed population structures for the lower levels, is to use the *ratio* method approach.

We can derive the lower level estimates for census years only – and to some extent from the Community Survey years too). These estimates are easily balanced to the provincial cohort component models.

To be more precise however, we break the population down using all of the characteristics present in both the top level (provincial) and lower level (census / CS) data. From here, we derive adjustment factors for each region and characteristic that, combined, will result in the sum of all the regions in a province being equal to the output from the provincial cohort component model.

The cohort-component model outputs population estimates per province, for each of the population groups, gender groups and five-year age groups. We use the Census estimates to derive the breakdowns for the low level populations, using the same characteristics. This results in a balancing equation which forces 128 different population characteristics (4 population groups by 16 age groups by 2 gender groups) for each region to balance to the provincial sum.

For the years between censuses, we interpolate the characteristic share for each characteristic. For projections beyond the latest census, we make an assumption about the growth of a region, typically taking a marginal decreasing change assumption. Final adjustments are made to the interpolated shares where necessary and where underlying data supports the changes.

We also take a proactive approach in consulting the latest literature on sub regional estimates, particularly the annual BMR reports on local municipal population estimates – and we continually sanity check our outputs on this level with our in house socio-economic models.

9. Multi-Model Method for Small Area Estimates

As is the case with local municipal estimates, there is not enough data to build cohort component models on a small area level. Once again, the major problem is migration between small areas. Even census data on this level has known shortcomings. We therefore introduce a number of additional data sets into the model at this point, which means that the major datasets for this portion of the work consist of the following:

- All available population censuses that were released on a low spatial level (1996, 2001 and 2011 are available at varying degrees of spatial granularity.)
- National Land Cover remote sensed satellite data (1995, 2000 and 2013 are available at the required resolution and national consistency.)

- Night time luminosity as provided by NOAA and processed by IHS. (This data is available for all years from 1992 to the present, with the later data available monthly.)
- Slope and elevation data (terrain slope is generally assumed to be stable over time, and is provided by various satellite data providers.)

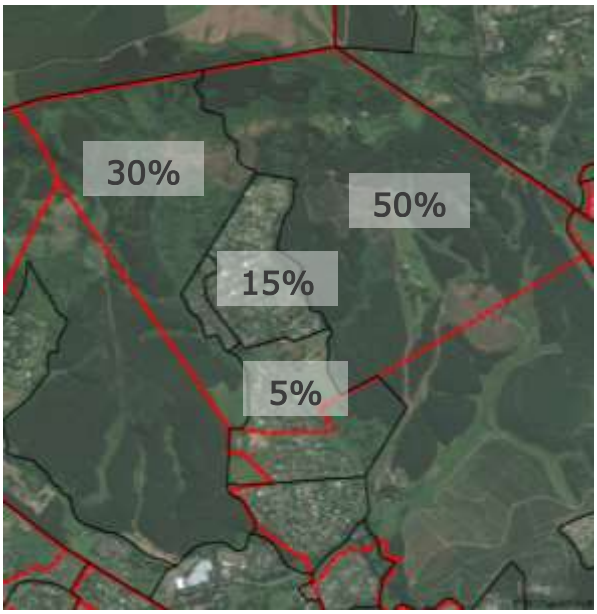
We do not make use of any official spatial development frameworks in this model. Although this is a potential future quality improvement, the data was not going to be suitable for this specific estimation.

Each of the above data sources has its own set of strengths and weaknesses. The land cover data, for example, has a very high resolution but is available at highly infrequent periods. The luminosity data has the exact opposite properties – its resolution is only *just* within a useful range, but it is available at a *monthly* time interval. We therefore combine these data sources in varying ways, calibrating each method such that it is both meets realistic sanity checks *and* that its output is similar to all of the other methods.

One consistent time series

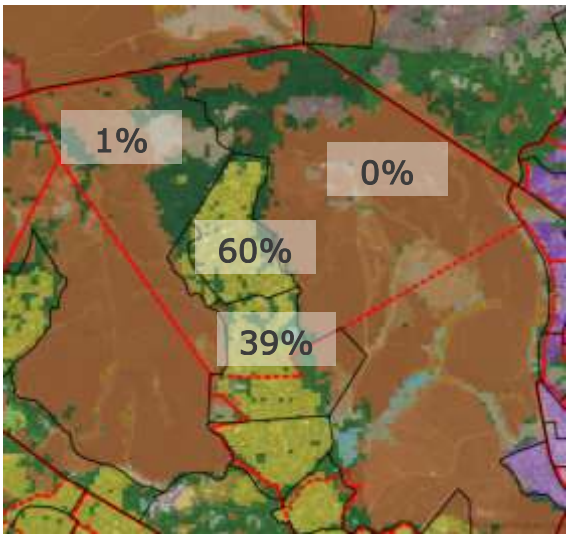
The first step to modelling population on this level is to create a consistent time series of the census data sets. The census data is a desirable target in that it provides a lot more detail than the other data sources; population by age and gender, proxy income brackets, unemployment and so on are all available from this data. However, its weakness is that the boundaries are not consistent from one census to the next.

This problem is typically solved by overlapping each of the data boundary sets with the target boundary set, and apportioning the target variable using a straight area overlap between the boundaries. In other words, if we wanted to get derive the total population from Census 1996 on Census 2011 boundaries, we would overlap the two boundary sets, and simply allocate the population from the 1996 boundary among the 2011 boundary on the basis of the area that the former overlaps the latter. The weakness to this approach is that it assumes the target variable is equally distributed across the input region – which in most cases is not a bad assumption.

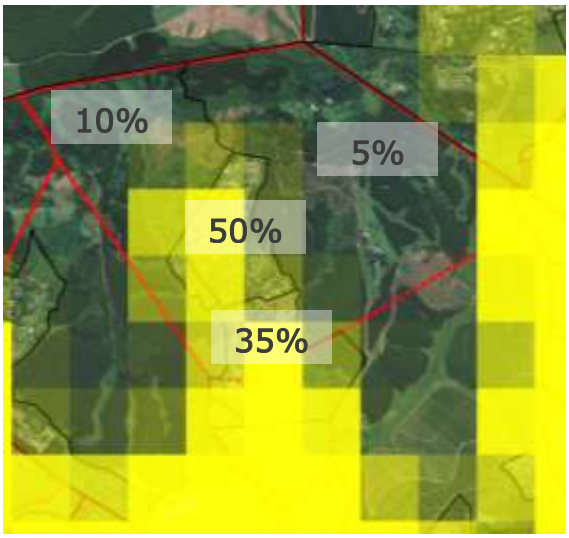


Area Overlap

However, in very specific – but important – cases, this assumption does not hold. In order to overcome this weakness, one can make use of either the national land cover data – which has fortunately been released for years very close to the census years – or the annual luminosity data from the given census year.



Land Cover Based Overlap



Luminosity Based Overlap

The above methods – combined in varying ways, with a bias toward the land cover-based data – provide a significantly better method of distributing the population for older censuses to the boundary set of newer censuses. However, this method is not entirely fool proof either. Although the land cover and luminosity data pick up total population distribution, these sources are not able to detect population group, age and gender breakdowns. Nonetheless, it is a vast improvement on the area overlap technique and is sufficient for the reporting requirements of this model.

Once balanced to the local municipal estimates from above, the output of this process is a single, consistent time series of population data from 1996 to 2011 by population group, age and gender.

Projections beyond the latest census

The consistent time series population data, alongside the higher level (local municipal) constraining data is useful in projecting the population beyond the latest census year. However, the variety of data sources and methods available lead us to develop four distinct models – briefly described below.

The **proxy cohort-component demographic** model is the first method that feeds into the final numbers. This approach attempts to build a model similar to a fundamental cohort-component model on a very low spatial level. This is possible because we have the population structure over

time, and can make some assumptions regarding births, deaths and migration for a specific region.

The advantage of this method is that it maintains the *age-gender-race-structure* of the population over time. It is also fairly stable over time, a feature driven by the fact that it is based on demographic fundamentals, and not on proxy data sources. The major downside to this method is that it is based upon historic data. Importantly, this means that it cannot detect new growth areas at all. Unsurprisingly, this method is also highly sensitive to the migration assumptions one makes on these low spatial levels.

The second model uses **luminosity change** as the driving proxy in detecting new growth areas. In this model, we calibrate the historic luminosity scores to accurately reflect the population change between known years (say 2011 and 2001). In other words, we use existing and known population data from the census to extract the range of luminosity scores that correlate well with population change.

Using the luminosity bands that have an observed correlation with population change, we calculate a luminosity-delta on the target boundaries – providing us with areas which are increasing in brightness, and therefore possibly have growing populations. The major advantage to this method is that luminosity data is readily available on a monthly basis from NOAA, is inexpensive and is fairly easy to work with. Currently, the data is released with an approximate two month lag, enabling us to detect very recent low level population change.

The downsides to this data however, are many. For starters, it is at a much lower resolution than required. For large rural areas this is not an issue, but for smaller areas it is not uncommon to find that luminosity scores overlap one or two small areas completely – making it impossible to determine which area is experiencing the luminosity delta. We can overcome this problem to some extent by interpolating to the required level, but this operation implies a number of additional assumptions of the source.

The other major problem with this data is that luminosity delta does not always imply population change. This is particularly true in rural areas which have undergone major electrification projects. Simply providing electricity to a community of people that already exists does not imply additional people moving to the area. On the contrary, the area may in fact be experiencing population decline and simultaneous electrification.

The third model takes a **land cover density** approach. Under this model, we compare the national land cover data with the known population by a given target boundary set. We then calculate coefficients for each of the land cover classes such that, by multiplying the regional share of land

cover by the calculated coefficient and summing all land cover data we end up with the total national population. These coefficients are then applied to the land cover shares of each small area layer, and balanced to the constraining local municipal populations.

This method has the advantage of using very high resolution source data, which is of a fairly good quality. However, it too suffers from many downsides. Primarily, this method breaks down fantastically in the presence of high variability across land use classes – such as that seen in metro areas where high rise buildings might lead to very large densities. Secondly, the data is not always entirely accurate, particularly when it comes to classing the differences between (say) industrial and commercial regions, or commercial and residential regions. Finally, the data is released at a very low frequency, with the implication that this model cannot stand alone without additional assumptions or source data.

The final method was a classical **cellular automaton model**, which attempts to take into account the way that urban areas spread across the terrain. This method attempts to combine all of the available data into a single model that operates on a 'cell' level – an arbitrarily low spatial level.

In particular, land use data from all the available land use surveys is used as a base dataset – distinguishing between urban non-residential, urban residential, rural natural vegetation and rural non natural vegetation. Each of those four land use classes are 'grown' annually according to a set of rules, roughly outlined as follows – bearing in mind that a cell will switch to a given class if the probability of switching to that class exceeds 50% in any given year:

1. If it is known that the cell will switch to a given land use class (say, because we have later land use data available) then give the cell a cumulative switching probability such that it will switch in the mid-year between the known data points.
2. If luminosity in the area increases from one year to the next, increase the probability of the cell switching to an urban classification – and vice versa if the luminosity decreases.
3. Increase the probability of the cell switching to each of its neighbouring cells (such that once all the cells around it are classified a certain way, the cell will switch in the next year to match those of its neighbours – unless it is known never to switch based on (1)
4. Never switch a cell to urban if it falls within an uninhabitable land class or in an area that has too much slope (based on terrain ruggedness.)

Although this method is able to pick up population change between known years very well, it has the possibility of growing way out of sanity bounds for very dense urban areas, and is particularly problematic in regions that are very small.

Finally, all of the above models are combined in a way that takes cognizance of each set of strengths and weaknesses. This takes the approach of a simple weighted average with some adjustments for known problems in each given model.