Technical Note – Population Back-Casting

1 Population Back-Casting

1.1 Introduction

The Saudi Arabian census of 2022 is the most meticulously carried out in the country's history. Census 2022 used administrative data to validate and cross check its accuracy and coverage. The Saudi Arabian population structure and demographic characteristics are best represented by Census 2022.

Comparing the results and the demographic structures of census 2022 with the population projections generated based on the census 2010, there were significant gaps that it is believed are combination of 2010 census carried overestimated errors and the use of assumptions in the past that did not take into account demographic changes that happened during the period between 2011 and 2021. This meant 2010 census population numbers and all population estimates based on it (years 2011-2021) had to be revised. The way forward for GASTAT was to conduct a population back-casting exercise.

Population back-casting is a widely used technique globally whereby statistics offices correct the annual population estimates issued in the years between census exercises in conjunction with the historical administrative data for deaths, births, and migrations while ensuring major socio-economic developments are correctly reflected in the data.

Population back-casting is the process of revising previous population estimates by applying statistical methods that takes into account the available administrative data to estimate related demographic indicators. The process was applied for each year from 2010 – 2021 at both national and regional, level, and also at the level of different cohorts by age, sex, region and nationality (Saudi or non-Saudi).

The administrative data used for the population back-casting processes as well as its characteristics and the relevant assumptions are described in the next part. Section XX details the methodologies, approaches and instruments that were employed. Section XX shows the major outcomes of the finalized population indicators over time. Section YY sets out the historical context of events that might have had an impact on the demographic development in the past decade.

1.2 Data and Model Parameters

This section introduces the administrative data used to implement the back-casting process and to validate the results. The administrative data, which were made available to GASTAT only in aggregated form, include data on the evolution of the population living in KSA (e.g., administrative births, deaths), as well as data on international migration flows, e.g., Trip in, Trip out, the data on the first entry (obtaining an employment visa and first-time entry into the nation), and administrative student and employee data. GASTAT was not able to validate the quality of the data on an individual record basis.

Table 1 summarizes the administrative data that is available, as was previously stated by data source, entity, frequency, and period:

Table 1

Data Source	Entity	Aggregated Data and Indicators	Frequency	Period
Administrative records	National Information Center (NIC)	Birth	Monthly	2010 - 2022
		Death	Monthly	2010 - 2022
		International flow (Trip in and out)	Monthly yearly	2010 - 2021 2011 - 2022
		First entry	Monthly	2011 - 2022
		Domestic workers for non-Saudis	Quarterly	2017 - 2022
	Ministry of Human resources and Social Development	Government sector workers	Quarterly	2017 - 2022
	General Organization of Social Insurances	Participants on the job who are subject to social insurance laws and regulations in the private and government sector, and those who stop participating in social insurance in the quarter according to gender, nationality, and reason for stoppage.	Quarterly	2017 - 2022
	Ministry of Education	Student totals by educational levels	Yearly	2014 - 2021

Administrative data was used as model parameters in the following manner:

Administrative Deaths ($\boldsymbol{D_t}$) used as a main parameter as the methodology requires death input.

Depending on the degree of analysis of the internal and international migration, the migration amounts immigration M_t^{in} and emigration M_t^{out} were composed of several components.

Internal Migration: On this subject, there is no administrative information. Later in this text, the techniques utilized to distribute internal migrations between regions overtime will be covered.

International migration: used as a main parameter to estimate net migration. It is crucial to distinguish between people entering or departing the country permanently for the purposes of tourism, hajj, or business, vs. people coming or leaving temporarily for the same purposes. The UN Statistics Division suggests that nations use one of two criteria when defining place of regular residence: (a) The location where a person has lived continuously for the majority of the past year (i.e., for at least six months and one day), excluding brief absences for vacations or work assignments, or where they intend to stay for at least six months; (b) The location where a person has lived continuously for at least the past 12 months, excluding brief absences for vacations or work assignments, or where they intend to stay for at least six months. However, to track immigrants by length of residence and emigrants by length of absence, the application of this criterion necessitates a high level of statistical complexity driven by the underlying data and accessibility to this data. In order to capture migration flows that are critical for population back-casting, residency status was used with the relevant monthly entries (Trip in) and exits (Trip out) being those of Saudis or foreigners with a residency status of resident non-Saudis and non-Saudis entering for the first time.

The administrative data that is available, as was previously stated, is a monthly travel data per year. Trip in and Trip out statistics were taken into consideration for the Saudi. The monthly first entry statistics were taken into consideration for the non-Saudi in addition to the monthly Trip in and Trip out.

Saudi migration from administrative statistics is calculated differently than non-Saudi migration. Statistical smoothing for Trip in and Trip out was used to estimate Saudi migration in order to address seasonal travel fluctiuations. While, non-Saudis migration, it is determined by using administrative Trip in, Trip out, first entry, and administrative education and employment statistics. More information about estimating migration is explained in the methodology section.

The system of migration statistics might be affected by uncertainties because some people with visitor visas may have overstayed the validity of their visas and stayed illegally in the nation. Additionally, some immigrants may have crossed the border without being detected by the official border control. However, it is assumed that these quantities are not significant and do not impact the accuracy of the calculated estimations of M_t^{in} and M_t^{out}.

1.3 Methodology

This section introduces the methodologies that has been taken into consideration to produce the back casting population projection:

1.3.1 Cohort Component Method

Based on the 2022 census data, the cohort component method is applied to estimate the total population size for each period that starts from 2010 to 2021 in a year-by-year approach. Besides the overall population size, the estimation also covers the number of males and females by yearly age and per nationality. The approach to get to the previous year's population number works by subtracting from the current year's population number the number of births, adding the number of deaths and subtracting the number of net migrants during the period that starts from previous midyear to current midyear. This is formalized in the following cohort component method equation:

$$P_t = P_{t+1} - B_t + D_t - M_t^{IN} + M_t^{OUT} \qquad (1)$$

Where: P_t is the previous year mid-year population value, P_{t+1} is the population base at time t+1, B_t is the number of births occurring between time t and t+1, D_t is the number of deaths occurring between time t and t+1, M_t^{IN} is the number of immigrants from the country during the period t to t+1, and M_t^{OUT} is the number of emigrants from the country during the period t to t+1. By limiting the population in year t to those over age 1 ($x \geq 1$), there is no need to explicitly consider the number of births (B_t) and formula (1) can be rewritten as:

$$P_{t-1} = P_t(x \geq 1) + D_t - M_t^{in} + M_t^{out} \qquad (2)$$

Where x stands for age. For all practical purposes, therefore, births can be omitted from the backward estimation equations.

The process needs to be applied year by year going back from 2022 to 2010 and is applied not only at the level of the aggregate national population, but also at the level of different cohorts by age, sex, region, and nationality (Saudi or non-Saudi).

1.3.2 Statistical Methodologies to Correct the Aggregated Administrative Data

This section introduces methodologies to statistically correct the aggregated administrative data before implementing them in the Cohort Component Method. The statistical methodologies are used to improve the available aggregated administrative data age structure and its totals.

The administrative data, Trip in and Trip out, that were received for this exercise are aggregated and contained errors in the age distribution. The error in the age distribution is coming from using one reference date to calculate the aggregated age. The methodology will be presented in the following section. Additionally, the administrative data, First Entry of Non-Saudis, is considered as well where the first entrants' age distribution is not accurate. For instance, there are more first entries in various age brackets (21, 22) than there were non-Saudis in the census at that time. It was possible to correct these errors by estimating the ages which are consistent with the population and administrative employees and student age distribution. The methodology will be covered as follows.

Using the administrative flow data on international migration, Trip in and Trip out, at the national level makes it difficult to distinguish between the migrants and short term travelers, and no information was available on who resided outside the Kingdom for more than 12 month\year over time. Additionally, no information was available on the regions from which these persons originate or on their destinations toward which they were moving. Later in this section, the techniques utilized to correct the net migration for Saudi and non-Saudi will be covered.

1.3.2.1 Estimating the Saudi migration age structure and totals:

This section introduces a method to address quality issues of aggregated administrative data on the migration of the Saudi population (Trip in and Trip out, age distribution).

The age of the received aggregated administrative data, Trip in and Trip out, was calculated by using Aug 2022 as a reference date for the whole traveler historical time series such as the population in 2011, their age reference date was Aug 2022. This error led to an elder population in 2011 than their actual age. As a result, the age of the traveler in a year other than 2022 is higher than their anticipated age by the gap of the same year and 2022. To correct the error in the received aggregated administrative data, Trip in and Trip out, the age distribution is recalculated using an age gap shifting that contrasts the expected reference date with the actual reference date.

After correcting the age profile, a statistical technique was used to estimate the Saudi migration totals. The technique that was used to smooth the Saudi Trip in and Trip out data from admin sources so that they conform more closely to expected patterns, assuming that the irregularities are caused by errors for example when certain administrations do manual recording of data, there is a higher chance of incorrect data entry. Moreover, it may appear due to having the short term traveler among the migration in the trip in and trip out data. As progress on improving administrative data is ongoing with the automation of the recording process. Unless significant historical events are known, the peculiarities of the data will be treated as errors and removed through data smoothing.

Data smoothing techniques may involve calculating averages using various measures of central tendency, calculating moving averages, aggregating data, or employing specific mathematical models – in this back-casting exercise, in order to solve the data quality issue with the Saudi migration data, a weighted moving average method was used to smooth historical migration data for Saudis only.

Weighted moving averages are a set of averages calculated from sequential segments of data points over a range of values, while assigning weight values to each of those data points and ensuring that the sum of those weights equals 1 (100%). It has the following formula:

$$M = \frac{\sum_{t=1}^{n} w_t v_t}{\sum_{t=1}^{n} w_t}$$

Where M is average value, v_t is actual value, w_t is weighting factor, and n is number of periods in the weighting group.

Extensive experimentation was conducted with varying combinations of window sizes and window types. The results were optimal for window size 5 with “triangular” window type, with weight values shown in Figure 1, for the migration variables (Trip in & Trip out) because they most effectively removed the noise while still capturing the patterns showing significant historical events.

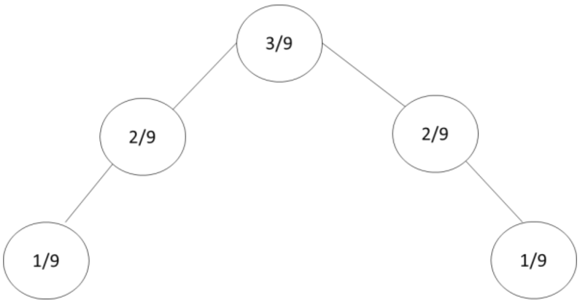

Figure 1 Weights for final experiment with window size 5 & window type triangular.

Figure 2 illustrates the effect of smoothing as part of the final experiment. It shows the results before and after data smoothing of the admin data variables for the Saudi male trip net (Trip in minus Trip out)

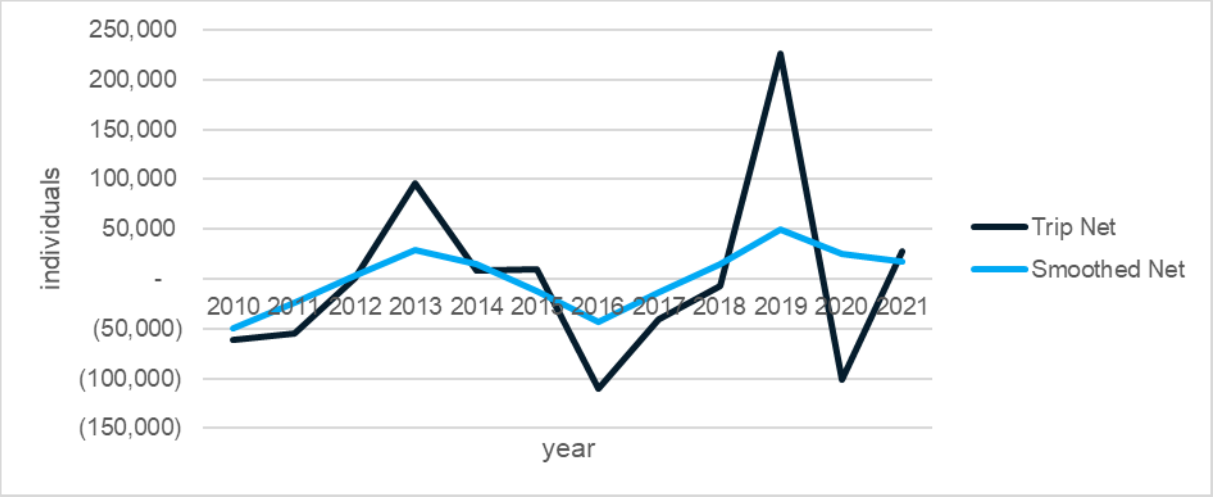

Figure 2 Saudi Males trip net before & after smoothing.

1.3.2.2 Estimating the Non-Saudi migration age structure and totals

This section introduces a method used to estimate the non-Saudi migration data i.e., the non-Saudi age structure and totals for the international flow. The non-Saudi international flow data divided into Entry (Trip in) and Exit (Trip out) of Resident and First Entry of Non-Saudis. The methodology includes correcting the age structure for the aggregated administrative data and totals. The methodology is to correct the non-Saudi age profile and total per year. As the correction of the next year cannot be conducted until the back-casting population base is formatted and estimated.

For the reason that the non-Saudi aggregated administrative Trip out totals exceed the Trip in for the whole time series, the Trip in is considered to be underestimated, the non-Saudi entry. Based upon, after communication with the data provider, it addressed the need of using the first entry data with the non-Saudi Trip in to estimate the migration. Based upon, Trip in and First Entry combined to be considered as M_t^{in} and Trip out was considered as M_t^{out}.

Furthermore, the non-Saudi data is required improving in the total as same as the age structure. The following main factors that it is believed they have impacted the quality of non-Saudi administrative international flow data: (a) The implementation of fingerprint in the border impacting the non-Saudi international flow data. (b) The misreporting of non-Saudi age as it has been noticed for First Entry of Non-Saudis. For instance, there are more first entries in various age brackets (21, 22) than there were non-Saudis in the census at that time. (c) Using aggregated data with misreporting age and total in addition to having an uncorrected aggregated data methodology. Those rise to the need of using several sources to estimate the non-Saudi migration age distribution and total.

Three solutions were considered in adjusting the age distribution in the most relevant age range. While they differ in the way that the first entries are redistributed by age, but they are based on the principle of consistent error minimization, applied partially, to a specific variable, rather than comprehensively, to the entire back-projection model. This principle dictates that the parameters of the distributions should be chosen in such a way so as to minimize the inconsistencies (in this case the inconsistencies between number of first entries and the total non-Saudi population in certain ages) resulting from a particular choice of parameters. The suggested solutions were applied to the age start at 21 to the upper limit age that is chosen in the interval (39,90). The chosen of the upper limit age is based on where the problem of age distribution seemed most pressing in the same year. The age distribution of children under age 21 seemed to be a lesser priority as the number of these children was quite small.

Based upon, the first solution is to estimate and allocate the first entry age distribution among those aged between 21 and the age upper limit in the interval (39,90) by using an existing distribution such as the population distribution, Trip in, and Trip out. Doing this exercise every year in the population back-casting insures the consistence between the net migration and population. Because of the non-Saudi data quality, the using of the most recent age distribution is considered as a based adding to that the correcting of the first entry age distribution. To determine the migration total, a comparison between the population back-casting result after implementing the corrected net migration with the labor force registration. The differences used to assume the underestimated migration. The key factor is the total employees cannot exceed the population for the age fifteen plus (15+), which means the employee are part of the population 15+. Based upon, the employment registration data play a main role in setting the assumptions for the population 15+ totals per age group to the non-Saudi migration data which mostly the area where the improvement was needed. Thus, to have a consistent backward population estimate with the employment statistics, it was thought that an assumption should be added to the Trip in or Trip out. The population numbers for the age groups of people under-19 years were validated using the student registration data. The assumption was that enrollment numbers by age group should not exceed the number of children as provided by the retroactive estimates of the population in the appropriate age group in the back casting process. This methodology is repeated every year to estimate a consistent migration totals and age structure by comparing the result with the available administrative student and employee data.

The second solution is a one-parameter (α) shift of the original distribution $M_t^{firstin}(x)$, i.e. first-time entrants of age x, according to an exponential transformation of the cumulative distribution of $M_t^{firstin}(x)$, normalized to a total of 1. The formula for this exponential age shift of the distribution is as follows:

$$M_t^{firstin(\alpha)}(x) = M_t^{firstin}(total)\{(\frac{\sum_{y=0}^{x} M_t^{firstin}(y)}{M_t^{firstin}(total)})^{\alpha} - (\frac{\sum_{y=0}^{x-1} M_t^{firstin}(y)}{M_t^{firstin}(total)})^{\alpha}\} \quad (3)$$

where $M_t^{firstin}(x)$ is the total number of first-time entrants and the summations. This for y=0 to x and x-1 refer to first-time entrants up to age x or x-1. This redistribution method reshapes the original distribution so that it skews more to the left (α<1) or to the right (α>1). The advantage of this procedure is that it preserves certain characteristics of the original distribution, especially the total number of migrants, thereby avoiding the need for a final normalization to obtain the correct total. However, when applying it to the data, it was found that the rather extreme adjustments that were required (α values much lower than 1) tended to cause an unrealistic heaping at the lowest ages of the distribution.

The third solution is to use the more conventional Gamma distribution:

$$M_t^{firstin(\alpha,\beta)}(x) = Cx^{\alpha}\exp{(-\beta x)} \quad (4)$$

Where x stands for age, α and β are shape parameters and C stands for the overall level. This distribution involves two (α and β), rather than one independent parameter and it requires a normalization at the end, to determine the value of C that will yield the correct number of migrants. Also, it does not preserve the characteristics of the original distribution to the same extent as (3). In particular, a disadvantage of the Gamma distribution is that it does not allow for the possibility of a bimodal distribution, which is a real possibility in the case of migration statistics.

Figure 3 compares the non-Saudi male first entry age distribution in 2016 with the new age distribution created using the mentioned methods. The first solution decreased the peak in the age 21–22 range and moved it to late twenty. The second solution, as moving to the left would result in unrealistic heaping at the youngest ages of the distribution. The second and third solution moved the male immigrants age distribution from being grouped around and age of 21/22 to having a peak in the mid-thirties. However, the norm of the first entry is having a peak in their twenty. Adjusting the parameters in the 2nd and 3rd solution to have more travelers in the later twenty led to an unrealistic age heaping. The 3rd solution is considered to be steadier moving the power to the left and it is closer to the first solution. Finding a reasonable distribution became more difficult, though, due to the presence of two parameters. The main challenge is 1): to define the potential adjustments to the data in terms of a minimal number of parameters α1, …. , αn (preferably not more than 50 and certainly not more than 100). In practice, this process usually involves some trial and error until all the quantities involved in have been satisfactorily specified. This process of trial and error in the specification of the model can be time-consuming. Based on the several experimentations, the first solution created a more stable migration total and age distribution and was considered to generate the non-Saudi population back-casting.

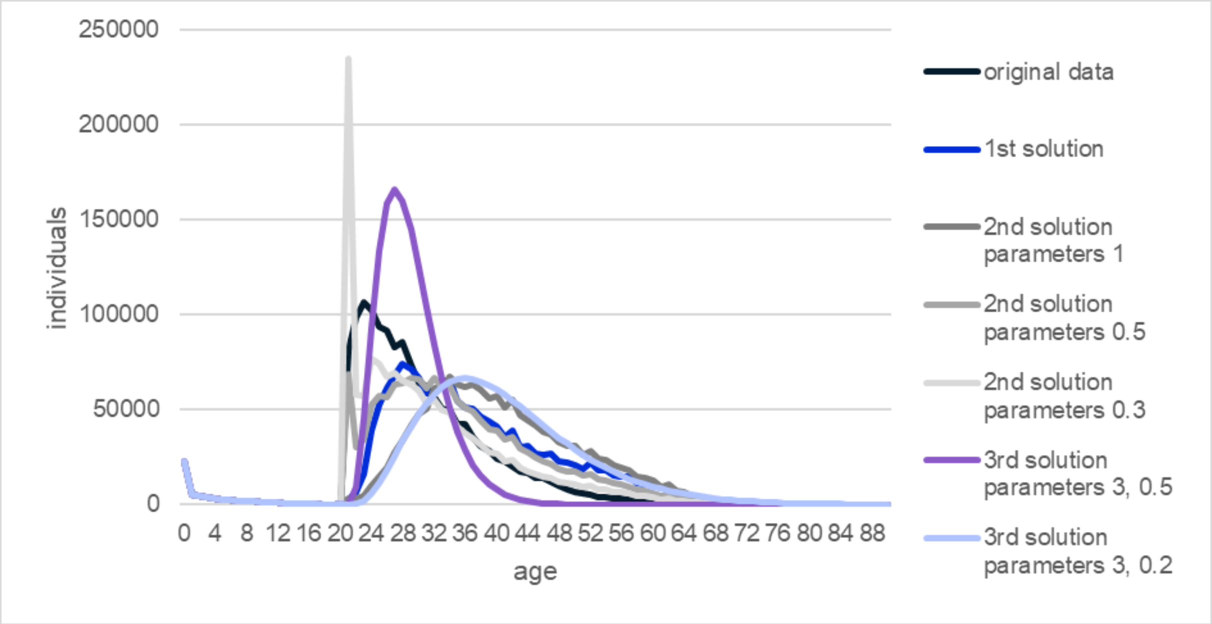

Figure 3 Non-Saudi male first entry example for several scenarios

1.3.3 Smoothing Single Years of Age

This section describes a smooth population age distribution methodology by using the moving average weights. It has the following formula:

$$S(x) = \frac{\sum_{i=-5}^{+5} w(i) \times P(x+i)}{\sum_{i=-5}^{+5} w(i)} \tag{5}$$

Where $P(x)$ is the population at age x. The considered age range is 5 with a default weight, $w(i)$ are:

Relative age	Sum	5	-4	-3	-2	-1	0	1	2	3	4	5
Trimmed weights	100%	0%	4%	8%	12%	16%	20%	16%	12%	8%	4%	0%

By using weighted moving averages, this analysis technique smooths a group by a single year of age. Although it uses different techniques than the mentioned before, it is very similar to the method used to smooth the administrative trip in and trip out data for the Saudi journey. It may make sense to attempt to smooth the population by a single year of age in situations where there is very strong age heaping on certain ages due to the administrative quality which affected the non-Saudi component more specifically non-Saudi migration.

The technique to even out the population's age spread by sex and age was only taken into account for non-Saudis in a particular year. With "trimmed weights," which use fewer age groups at the beginning and end of the age distribution.

The smoothing was put into place in 2018 and it is thought that administrative data beyond that point is less accurate than the most recent data, where the most recent data estimation is closer to being dependable. As the historical administrative data regarding non-Saudi birth, death, and migration improved adding to than the employment data and educational data. The population projection as a whole would advance.

1.3.4 Transition of regional shares during the 2010 – 2021 back-projection

This section introduces the regional share transition methodology for the period that starts from 2010 to 2021. The methodology is implemented after the population back-casting national level is generated. Its result drives the regional disaggregation process i.e., it estimates the regional total per nationality per gender where the disaggregation possess estimates the regional total by age group per nationality per gender, more information about the disaggregation possesses in the next section.

The need of determined the regional share for Saudi and non-Saudi is to estimate the internal migration changes in the regions over time. While only the initial internal migration result is available from census 2022 and it was collected for changing the place of residence in the past five years. Moreover, the past five years have high demographic dynamic changes that impacted the non-Saudi; the high number of non-Saudi emigration, it is believed to have an impact in the collected internal migration. Additionality, there is no data regarding the internal migration available from registration for the period between 2010 and 2022. Based upon, it is necessary to do the top-bottom technique to estimate the regions because of the internal migration quality and no regional data for international migration. The methodology is considered to be an indirect estimation of the internal migration over time.

There are two approaches to estimate the regional share transition are taken in consideration. The first approach uses the regional distribution from the 2010 census and 2022 census as a basis then estimates weights over time to generate the regional distribution in between. The regional division result per nationality of the first approach led to inconsistency in the results regarding the non-Saudi trend over time. An alternative solution is needed to be considered in order to overcome the irregularity. The alternative solution, second approach, uses instead the non-Saudi regional share per gender from the 2010 census and 2022 census as a basis then estimates weights over time to generate the non-Saudi distribution in between. The second approach generates more consistence result than the first approach and is considered in generating the population back-casting regional result. The following illustrates the two approaches in details.

The first approach as in the national level methodological approach, the evolution of the regional population starts out with the census populations of the regions in 2022 and back-projected regional populations in 2010. The latter were computed by applying the distribution of the regional populations according to the 2010 census to the back-projected national population. In other words, the regional populations in 2010 that were used for the regional disaggregation have the same distribution as in the 2010 census, but their total does not equal the total of the 2010 census, but the back-projected national population of 2010.

In the intermediate years 2011-2021 the computation starts out with the addition, i.e., the 2011 population of the region R in 2011 will be the 2010 population plus the number of estimated births occurred since 2010 per region. Adding these numbers over the regions results in a preliminary national total that considers births, but not deaths and migration. (The regional birth distribution in the back-projection was taken to be the same as in the 2022 census because the vast majority of births are among the Saudi population and because of their regional distribution being reasonably consistent.) This preliminary number can be compared with the back-projected national total where the difference between the two needs to be distributed among the regions. Because migration is the dominant factor in this difference, it is distributed according to the census 2022 initial internal migration. In 2010, it was presumed that the distribution of in-migrants was the same as the relative population distribution of the regions. The linear interpolation is used to guarantee that the overall total of the interpolated percentages will be 100%. The use of a forward and backward procedure for the distribution by region of the differences between the back-projected national total and the preliminary total derived previously (here to be referred to as national surplus).

In the forward procedure, in order to obtain the 2011 regional populations, first regional births are added to the 2010 population, as described above, and then the regional share of national surplus is added according to the regional percentage of 2010 in the distribution table. Starting with these adjusted regional populations, the process is then repeated for subsequent years, i.e., adding the births of year y to the initial population of year y and then prorating the national surplus according to the percentages for year y in the distribution table. The backward procedure operates in the opposite direction. It starts in 2022 and derives the 2021 population by subtracting the 2021 births from the 2022 populations and then prorating the national surplus according to the percentages in the distribution table for 2021. It is then applied successively to the previous years. The results are two time series for the regional populations in 2011-2021 which do not necessarily coincide. In order to obtain the final time series, a weighted average of the forward and backward time series was used. In the final result for 2011, a weight of 10/11 was applied to the forward series and 1/11 to the backward series. In 2012, a weight of 9/11 was applied to the forward series and 2/11 to the backward series, and so forth, up to 2021, when a weight of 1/11 was applied to the forward series and 10/11 to the backward series. The procedure was applied first for the total population per region. Based on the lifetime migration analysis, for more than 80% of the Saudis the current region of residence matches the region of birth which indicates that the stability in Saudi mobility over time. Based on that the Saudi total per region were projected by using the 2022 census going backward. Then the Saudi total was subtracted from the total of each region to obtain the regional total for non-Saudis.

The second approach as in the national level methodological approach, the evolution of the regional population starts out with the census populations of the regions in 2022 and the back-projected regional populations in 2010. The latter were computed by applying the non-Saudi per gender shares of the regional populations according to the 2010 census to the non-Saudi back-projected national population in 2010. In other words, the non-Saudi regional populations in 2010 that is used for the regional disaggregation have the same share as in the 2010 census, but their total does not equal the non-Saudi total of the 2010 census, but the non-Saudi back-projected national population of 2010. The Non-Saudi regional shares per gender over time are deducted using a linear interpolation between the regional population share in census 2010 and census 2022. In order to obtain the final time series, a weighted average of the census 2010 regional share and census 2022 regional share were used. See Figure 4

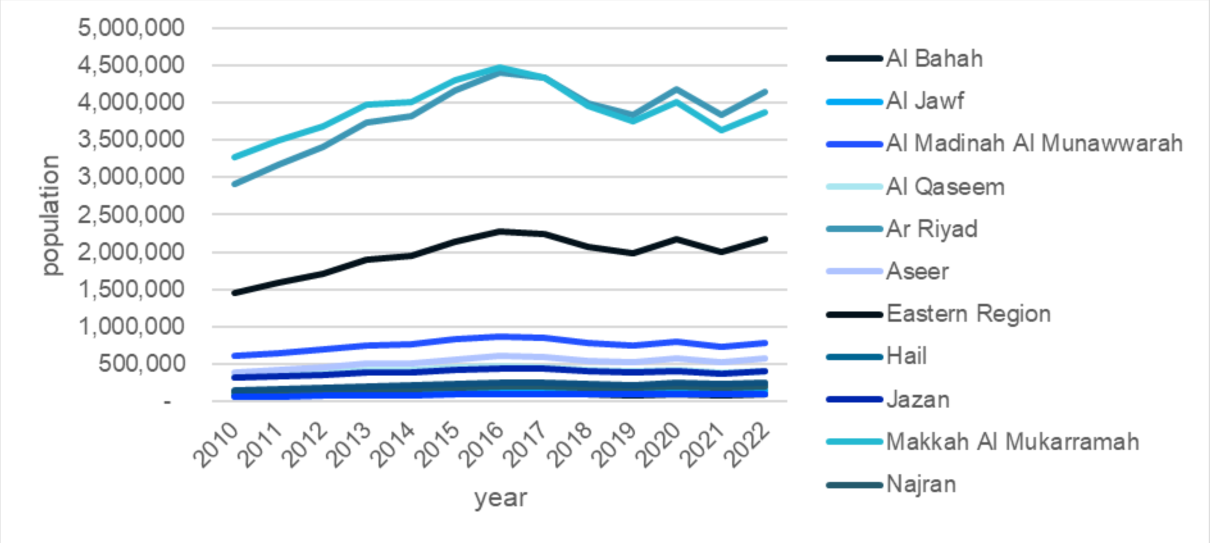

Figure 4 Non-Saudi regional distribution over time by total

The Saudi is generated as the first approach due to Saudi mobility stability over time when comparing the place of birth with the current residency region. See Figure 5

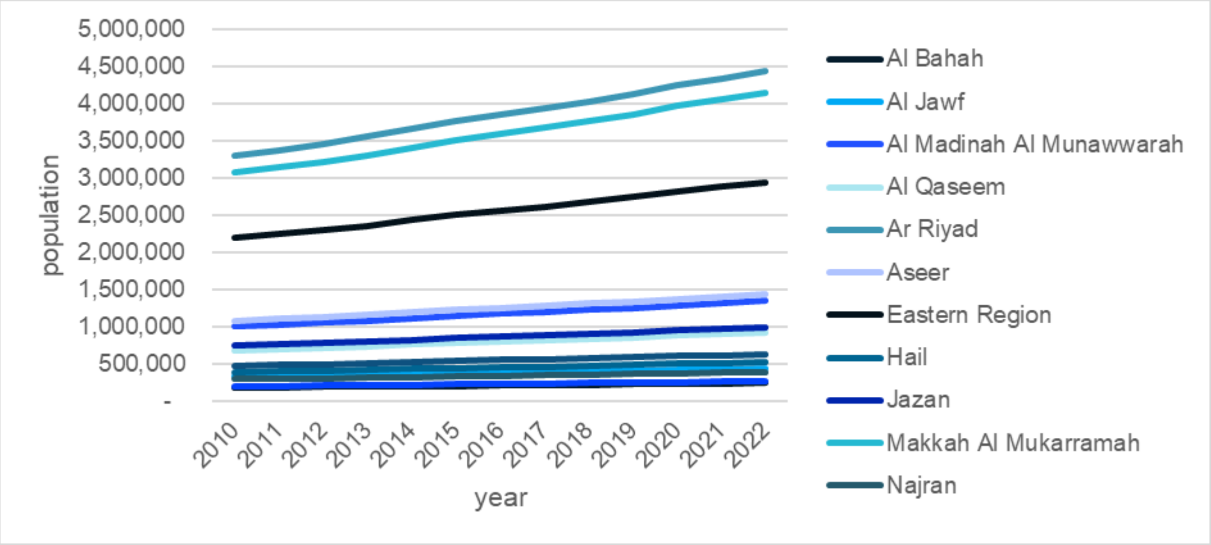

Figure 5 Saudi regional distribution over time by total

Thus, the Saudi and non-Saudi results are combined to generate the regional totals. Consequently, the regional distributions of the 2010 projected regional distribution for the back-projected population, is slightly different than the 2010 census. It is thought that the regional transition over time for the second approach reflects the consistency of the occurrences overtime. For instance, the increased density of the Riyadh region is reflected after the Vision 2030 implementation in 2016.

1.3.5 The disaggregation process for the regional level

This section introduces the disaggregation process for the regional level after generating the population back-casting national level. In the previous section, the regional total per nationality per gender were estimated. This section introduces a method that generates the regional total by the yearly age by nationality, by gender, which is defined as a disaggregation process for the regional level. In the same way as in the national level methodological approach, the evolution of the regional population starts out with the census populations of the regions in 2022 and back projected to 2010. The methodology is called the cohort-ratio method where it is based on regular cohort survival ratios, its regional totals are estimated by using the regional share transition methodology results. Concretely, the method can be described as follows:

Age	National level			Regional level		
	Year y	Year y-1		Year y	Year y-1	Attributed y-1
……						
x-1	$P_N(x-1,y)$	$P_N(x-1,y-1)$		$P_R(x-1,y)$	Unknown	$P_R(x,y)*P_N(x-1,y-1)/P_N(x,y)$
X	$P_N(x,y)$	$P_N(x,y-1)$		$P_R(x,y)$	Unknown	$P_R(x+1,y)*P_N(x,y-1)/P_N(x+1,y)$
x+1	$P_N(x+1,y)$	$P_N(x+1,y-1)$		$P_R(x+1,y)$	Unknown	$P_R(x+2,y)*P_N(x+1,y-1)/P_N(x+2,y)$
x+2	$P_N(x+2,y)$	$P_N(x+2,y-1)$		$P_R(x+2,y)$	Unknown	
……						

In other words, the cohort survival ratios at the national level (which are known from the national cohort-component back-projection) are attributed to each of the regions and multiplied by the regional populations of the appropriate ages in year y. This guarantees that the cohort survival ratios in each of the regions will be plausible (namely, equal to the national ratios). However, it does not guarantee that the results will be consistent with the regional totals or with the national totals by age group. In order to force this consistency, a bi-proportional adjustment has to be applied, which can be described as follows:

Age	Region 1	Region 2	Region 3	……..	Region 13	National total
……..						
x-2	P_1(x-2,y-1)	P_2(x-2,y-1)	P_3(x-2,y-1)		P_{13}(x-2,y-1)	P_N(x-2,y-1)
x-1	P_1(x-1,y-1)	P_2(x-1,y-1)	P_3(x-1,y-1)		P_{13}(x-1,y-1)	P_N(x-1,y-1)
x	P_1(x,y-1)	P_2(x,y-1)	P_3(x,y-1)		P_{13}(x,y-1)	P_N(x,y-1)
x+1	P_1(x+1,y-1)	P_2(x+1,y-1)	P_3(x+1,y-1)		P_{13}(x+1,y-1)	P_N(x+1,y-1)
x+2	P_1(x+2,y-1)	P_2(x+2,y-1)	P_3(x+2,y-1)		P_{13}(x+2,y-1)	P_N(x+2,y-1)
…….						
Regional total	P_1(tot,y-1)	P_2(tot,y-1)	P_3(tot,y-1)		P_{13}(tot,y-1)	P_N(tot,y-1)

- In the first step, the regional populations in each line are adjusted proportionally, so that they will sum to the correct national totals P_N(x-2,y-1), P_N(x-1,y-1), etc.
- In the second step, the regional populations in each column are adjusted proportionally, so that they will sum to the correct regional totals P_1(tot,y-1), P_2(tot,y-1), etc., which were obtained independently.

Due to the fraction multiplication in each iteration, which results in the sum of all regions not equaling the sum of all ages at the national level, the data will typically still be inconsistent after this initial stage. However, the process is repeated as many times as necessary (usually not more than 5 iterations) until the table is consistent both in the horizontal and the vertical direction. In the final iteration procedure, converting the fractional numbers results into entire numbers that maintains the horizontal and vertical consistency of the results.

1.4 Validation

This section introduces the administrative data that are used to validate the population back-casting results and their usage. The retro-projections are contrasted with the administrative data includes vital statistics and data on enrollment and labor force, in order to evaluate their reasonableness.

Administrative Births (B_t):

Used as parameters: The population back-casting methodology does not require birth input; however, it is essential for the validating the population back-casting.

Validation methodology: the validation is conducted by comparing the population back-casting age zero results with the existing administrative birth data per nationality per gender. For the administrative birth is calculated by using newborn residence status (Saudi or non-Saudi) per gender where the place of birth is inside KSA. For the mother, the assumptions include multiple mother nationalities (Saudi and non-Saudi) with a known mother age. The validation is done based on the national level because the geographical distribution of births is not documented in the available historical administrative birth.

Validation result: the population back-casting age zero per nationality per gender is higher than the administrative birth for the period 2012 – 2021.

Administrative Deaths (D_t):

Used as parameters: The population back-casting methodology requires death input.

Validation methodology: the validation is conducted by calculated mortality indicators such as life expectancy at birth e(0) and death completeness score. The validation is performed for Saudis only because the Saudi population is considered as stationary population and the results can be detected and generate consistent behavior. The considered data is the Saudi administrative death and the Saudi population back-casting result. The Saudi's death life expectancy at birth is calculated by using the life tables. The Saudi's death completeness score is calculated by using the Brass Growth Balance Method. The validation is done based on the national level because the geographical distribution of deaths is not documented in the available historical administrative death.

Validation result: the Saudi life expectancy at Birth utilizing administrative deaths and the population outcome of back-casting result over time is consistent as the life expectancy at birth is improving over time with the exception of the year reflecting COVID-19. The Saudi's death completeness score results illustrate that the Saudi male deaths appear to have high coverage, whereas Saudi female deaths appear to be under-covered.

School Enrollment:

Used as parameters: The population back-casting methodology does not require student totals input.

Validation methodology: the validation is conducted by comparing student totals per school level with the respective age group of the population back-casting result. School level in Saudi Arabia is divided into three levels: elementary education (six years), followed by intermediate education (three years), and secondary education (three years). The comparison is done by calculating the total differences between total students per school level and the related age and gross enrolment rate.

Validation result: for the Saudi, the average student total difference between the population and total student per educational level is around 15K where the average gross enrollment rate is 101%. For non-Saudi, the average student total difference between the population and total student per educational level is around 23K where the average gross enrollment rate is 89%.

Labor Force Registration:

Used as parameters: The population back-casting methodology does not require employee input.

Validation methodology: the validation is conducted by comparing the differences between the total of population back-casting age fifteen and above with total administrative employee in Saudi Arabia per nationality per gender per age group.

Validation result: the total of the population back-casting age fifteen and above is higher than the employee with respect to the nationality per gender, this step insures the reliability of the population back-casting result and the available historical administrative employment data. The employment share per nationality per gender per age group of the population over time is consistent.

1.5 **Used Tools**

Python software was used to automate the cohort component approach, cohort ratio method, and a bi-proportional adjustment methodology. Additionally, Python software was used for the analysis, preparation, and improvement of the administrative data. The advantage of automating back-casting is that it speeds up results, which prompted analysis of the results and scenario testing.

1.6 **Results**

The section illustrates the results of the population back-casting projection 2010-2021 and its comparison with the population of census 2022.

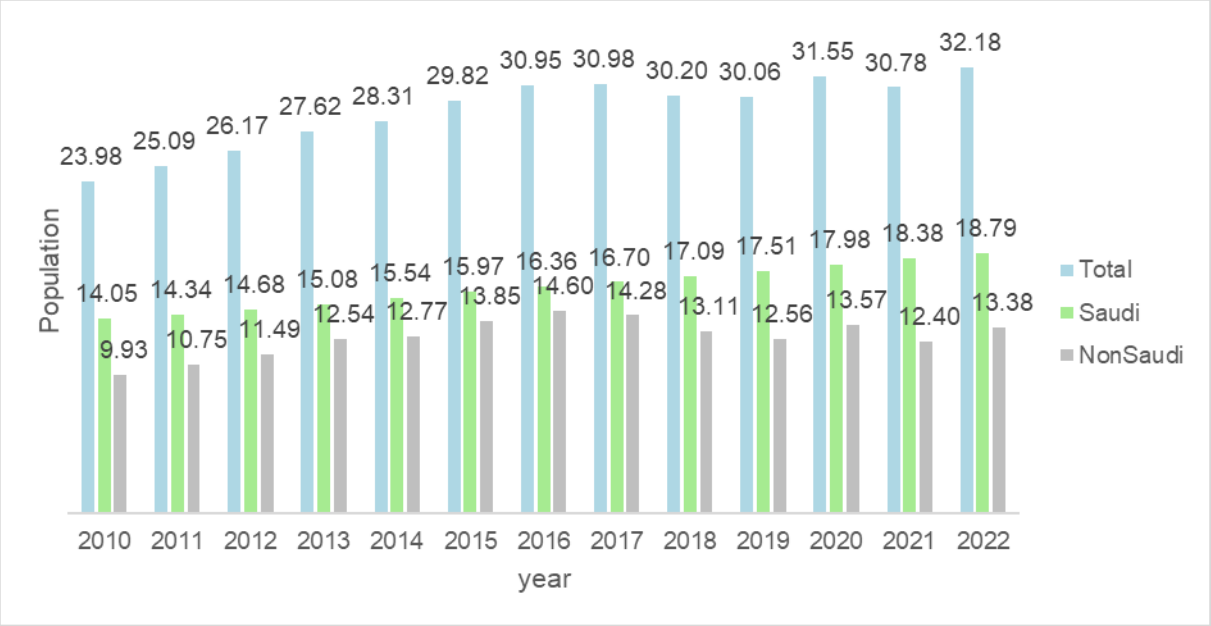

Figure 6 Population total per nationality over time (unit in million)

The total population in KSA amounted to 23.98 million in 2010 (mid-year), according to the population back-casting projection. The population increased by 34.18% in Census 2022 when the total population stood at 32.18 million compared to mid-2010. The main driving factor of the growth in the overall population over the twelve-year period is first the growth in the Saudi population by 33.79%, the main source for the Saudi growth is the Saudi natural reproduction (births minus deaths). Second, the growth in non-Saudi population by 34.74% during the same period, the main source for the non-Saudi growth is the non-Saudi immigration. See Figure 6

After the population numbers increased in KSA from 23.98 million in 2010 to 30.95 million in 2016, the population at midyear 2017, reach 30.98 million, saw a slowed growth compared to the previous year. However, the mid-years 2018 and 2019 saw a decline in the overall population numbers 30.20 million, and 30.06 million consequently. While the natural reproduction (births minus deaths) was positive for both Saudis and non-Saudis in 2017/18/19, the decline in the overall population is mainly related to the increase of non-Saudi emigration after the collection of financial compensation for escorts and escorts of expatriate workers began. After, the population numbers increased in KSA from 30.06 million in 2019 to 31.55 million in mid-2020, the mid-2021 saw a decline in the overall population numbers to 30.78 million. While the natural reproduction (births minus deaths) was positive for both Saudis and non-Saudis in 2021, the decline in the overall population is mainly related to non-Saudi leaving the Kingdom during the COVID-19 pandemic. Next, the population numbers increased in KSA from 30.78 million in mid-2021 to 32.18 million in the 2022 Census, the 1.4 million growths between mid-year 2021 and 2022 census is a result of increasing the total of the non-Saudi immigration, and birth correspond to a post COVID-19 pandemic. See Figure 6

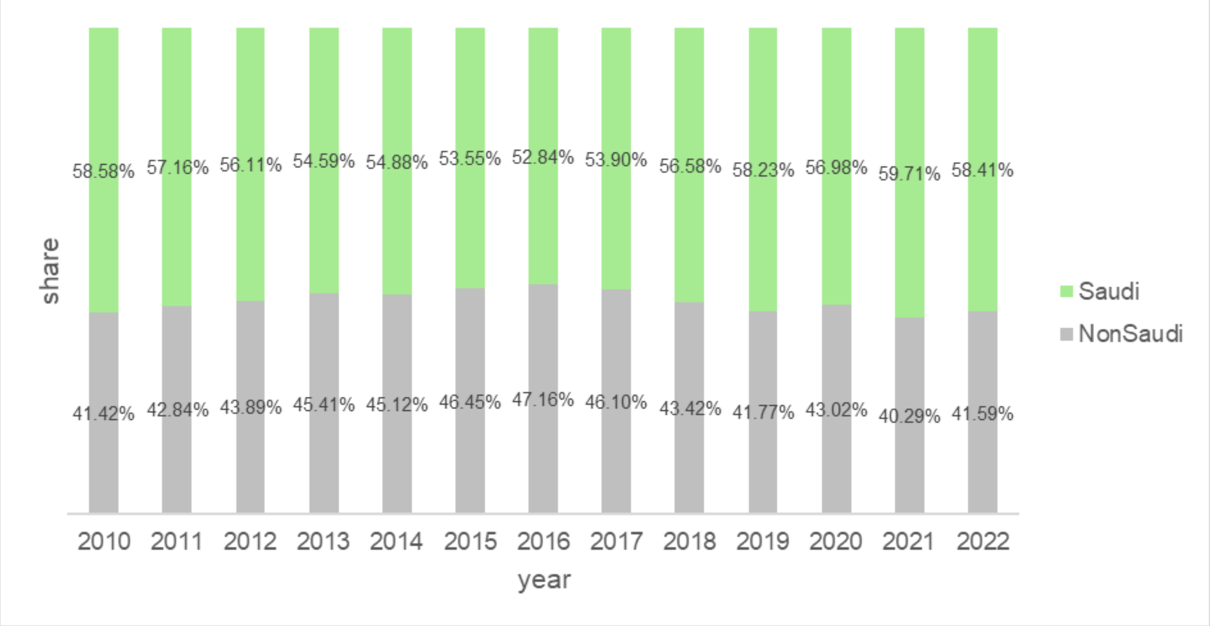

Figure 7 Population Share per nationality (Saudi and non-Saudi) over time.

In 2010, the Saudi population accounted for 58.58% of the total population, with non-Saudis at 41.42%. While in 2022, the Saudi population accounted for 58.41% of the total population, with non-Saudis at 41.59%. The Saudi population share starts to decrease from 2010 to 2017 as the non-Saudi immigration increase. However, from 2017 to 2019, the Saudi share saw an increase due to non-Saudi emigration. Based upon, the non-Saudi migration has an impact in non-Saudi share of the population which is in the inverse direction of the Saudi. See Figure 7

In 2010, Males accounted for 59.31% of the total population, or 14.22 million individuals, females made up 40.69% of the total population, with a population of 9.76 million females. For a gender share comparison for the combined nationality, the male share continues increasing over time for the whole time series. In 2022, Males accounted for 61.16% of the total population, or 19.68 million individuals, females made up 38.84% of the total population, while considering a population of 12.50 million females. For a Saudi nationality per gender, in 2010, Males accounted for 50.31% of the Saudi population, or 7.07 million individuals, females made up 49.69% of the Saudi population, with a population of 6.98 million females. For a gender share comparison for the Saudi nationality, Saudi female share slightly increased over time for the whole time series than the Saudi male. In 2022, Males accounted for 50.20% of the total population, or 9.43 million individuals, females made up 49.80% of the total population, while considering a population of 9.36 million females. For a non-Saudi nationality per gender, in

2010, Males accounted for 72.05% of the non-Saudi population, or 7.16 million individuals, females made up 27.95% of the non-Saudi population, with a population of 2.78 million females. For a gender share comparison for the non-Saudi nationality, non-Saudi male share continues increasing over time for the whole time series. In 2022, Males accounted for 76.55% of the total population, or 10.24 million individuals, females made up 23.45% of the total population, while considering a population of 3.14/ million females.

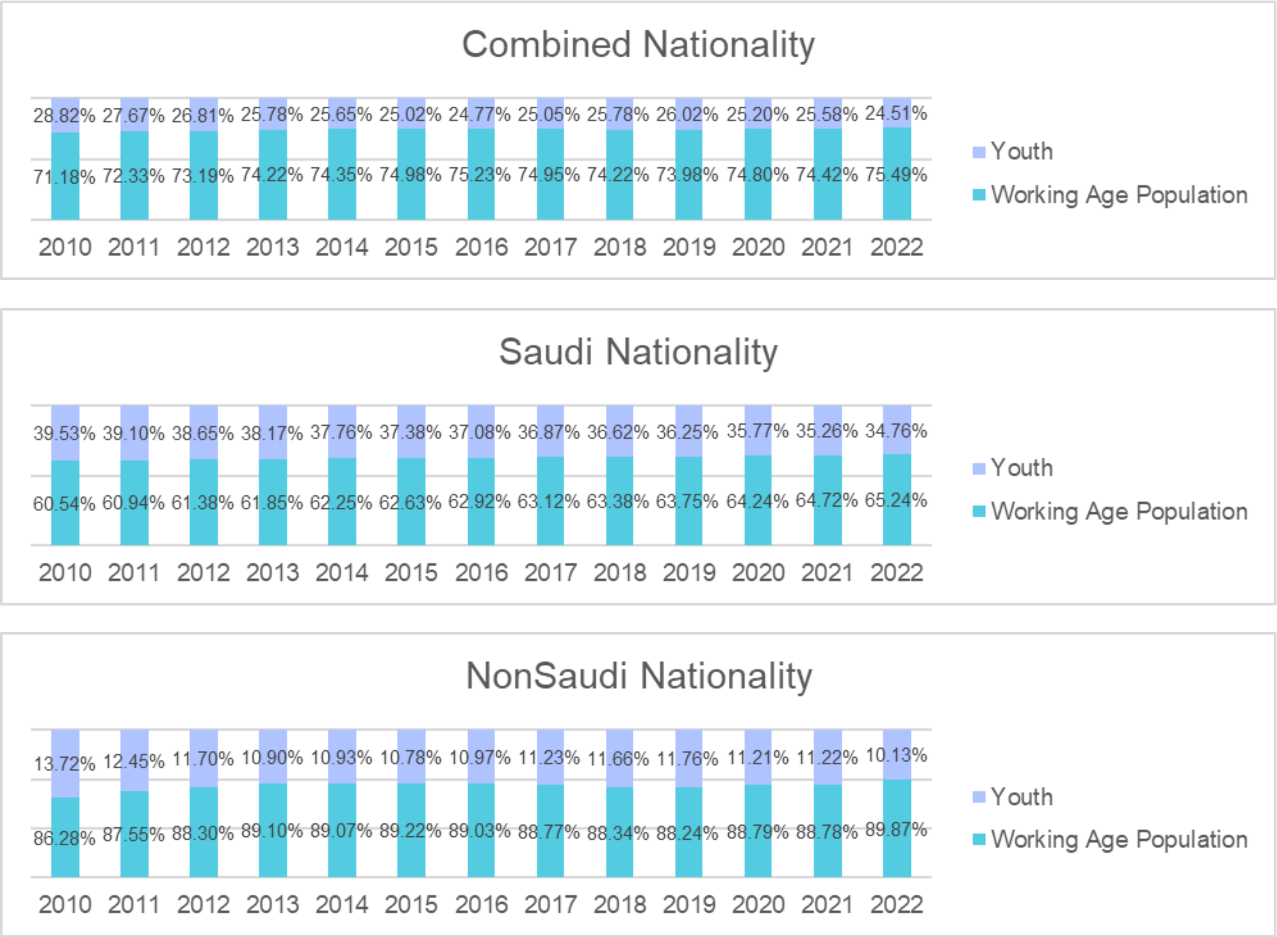

Figure 8 Population Share per nationality (Saudi and non-Saudi) per categories (Youth and Working age) over time.

The total of working population, population age fifteen and above, has increased by 42.31% in Census 2022 compared to mid-2010. While, the youth population total, population age less than fifteen years old, has increased by 14.11%. In 2010, the working age counted for 71.18% of the total population, with a youth population at 28.82%. While, in Census 2022, The working age counted for 75.49% of the total population, with a youth population at 24.51%. For the same period, Saudi and non-Saudi working ages have increased faster than youth-age. See Figure 7

2 Demographic related indicators

2.1 Introduction

To correlate the historical population projection with the population counted in the census of 2022, a population back-casting exercise must be carried out. The indicators that employed the population as a primary component in their methodology are thought to have been altered as a result of the population back-casting update of the population total per nationality, per gender, and per age in the period from 2010 to 2021.

Indicators of health related to mortality and fertility that are impacted by population back-casting are shown in this report. The study outlines the limitations of the historical birth and death data for the updated health indicators. The methodologies for the relevant indicators are described in the next section. The new indicators are then represented in the result section, along with a comparison to the ones that had previously been released.

2.2 Data and Limitation

This section introduces the data used to calculate the health indicators and their limitation.

Population: For the indicators of mortality and fertility, the population back-casting is taken into account.

Birth: The birth input in the fertility indicators is calculated by backcasting the population's age zero in the same way that the administrative births in the relevant year are distributed by mother age.

Death: The administrative death distributes by age group for the 12 months before the population reference date are used to estimate the death input in the mortality indicators.

Limitations on Birth and Death: Access to Individual level is not available for the administrative birth and death, which are presented only at aggregated level. Additionally, it is unfeasible to estimate the health indicators based on the health cluster nor regional level because historical administrative data on geographic distribution is not accessible. Furthermore, the regulatory structure, which forbids non-Saudis from remaining after the working period, is thought to have affected the historical administrative birth and death data. Based upon, some indicators are calculated solely for Saudi.

2.3 Methodology

The fertility and mortality indicators that are being revised are described in this section, along with their methodologies:

Fertility indicators

Age Specific Fertility Rate

The age specific fertility rate ASFR methodology is represented in the following:

$$ASFR_x = \frac{B_x}{Female_x} \times 1000$$

Where B_x is the estimated birth for the 12 months previous to the population reference date, $Female_x$ is the female population, x is the age group where x from 15 to 49.

Total Fertility Rate

The total fertility rate TFR methodology is represented in the following:

$$TFR = 5 \times \sum_{x=15-19}^{45-49} ASFR_x$$

Mortality indicators

Age Specific Death Rate

The age specific death rate ASDR methodology is represented in the following:

$$ASDR_x = \frac{D_x}{P_x} \times 1000$$

Where D_x is the death for the 12 months previous to the population reference date, P_x is the population, x is the age group.

Life Expectancy at Birth E(0)

The Life expectancy at Birth E(0) is estimated by using the life table where Coale-Demeny Model (East) was considered. The life table contains several columns, each with a unique interpretation. See Table 1 for life table notations:

Table 1

Notation	Definition	Equation
x to $x+n$	The period of life between two exact ages. Here x indicates the starting point for an age interval and n is the interval length.	
q_x	The proportion of persons alive at the beginning of each age interval who die before reaching the end of the age interval.	$q_x = \frac{2 * n * d_r}{2 + n * d_r}$
d_r	The age specific death rate	
p_x	The proportion of persons alive at the beginning of each interval who survives over the age interval.	$p_x = 1 - q_x$
l_x	Of the starting number of newborns in the life table [called the radix of the life table set at 100,000]. The number of people living at the beginning of each age interval.	$l_x = l_{x-n} * p_{x-n}$
d_x	The number of persons in the cohort who die in the age interval x to $x+n$.	$d_x = l_x - l_{x+n}$
L_x	Number of years of life lived by the cohort within the indicate age interval x to $x+n$ or person-years of life in the age interval.	$L_x = n * l_{x+n} + d_x * a_x$
a_x	The average of time lived in interval x to $x+n$ by those dying in the interval.	
T_x	Total person-years of life contributed by the cohort after attaining age x.	$T_x = T_{x+n} + L_x$
e_x	The expectation of life at any given age i.e., the average number of years of life remaining for a person alive at the beginning of age interval x.	$e_x = \frac{T_x}{l_x}$

Table 2 gives an ordinary life table where the columns from (2) to (7) are calculated by the mentioned equations in the previous notation table:

Table 2

Age interval x **to** $x+n$	**Year** n	**Population**	**Number of deaths**	**(2)** q_x	**(3)** l_x	**(4)** d_x	**(5)** L_x	**(6)** T_x	**(7)** e_x
0	1				100,000				Life expectancy at birth
1	4								
5	5								
10	5								
⋮	⋮								
60	5								
65	5								
70	5								
75+	+								

2.4 Result

This section describes the indicators' findings for fertility and mortality and contrasts them with those that had previously been made public.

Figure 1 displays the expected Saudi birth distribution by mother age for the years 2011 to 2021. As a Saudi mother's age changes with time, her birth behavior switches to delaying the birth.

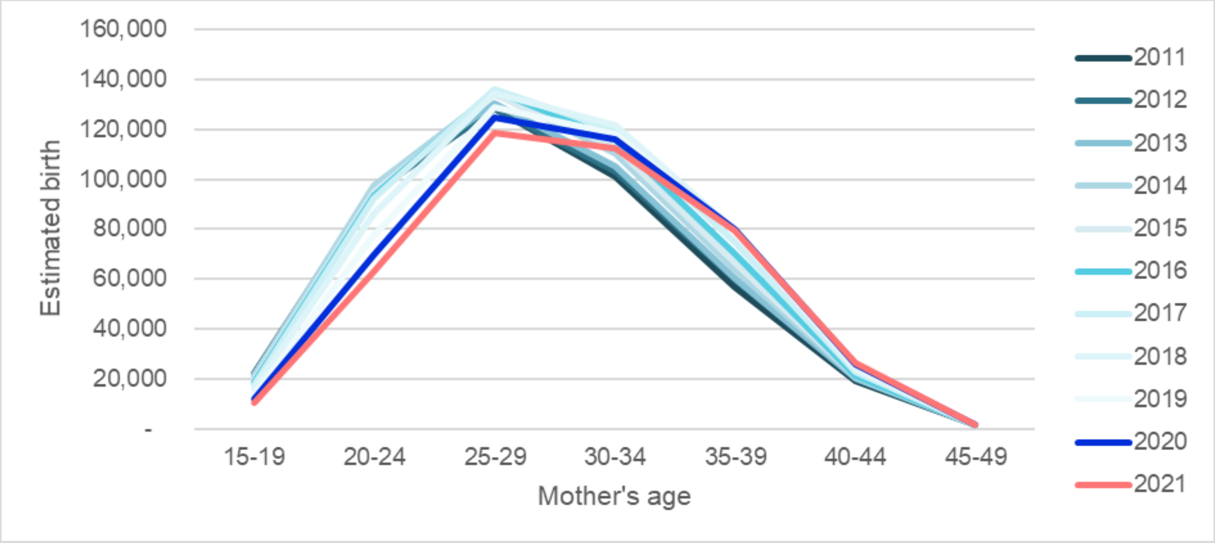

Figure 1: Estimated Saudi birth distribution by mother age over time.

Figure 2 compares the 2018 Saudi revised ASFR to the 2018 Saudi ASFR from the Household Health Survey. Comparatively, the Saudi revised ASFR predicts a higher overall birth rate.

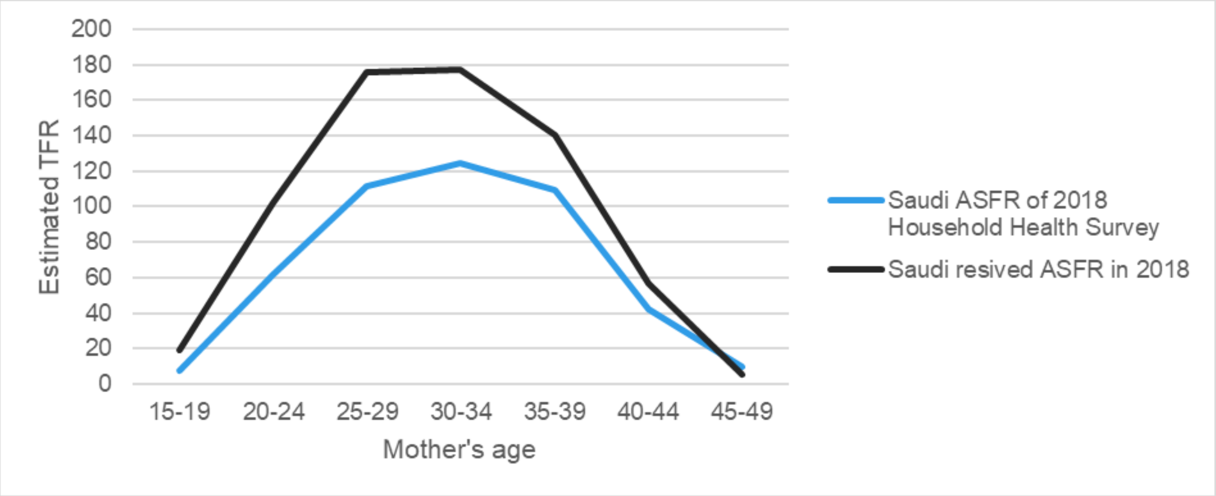

Figure 2: Comparison between the 2018 Saudi revised ASFR verses the 2018 Saudi ASFR from the Household Health Survey.

A further finding from the 2018 Household Health Survey revealed that there were 2.33 children per Saudi woman. But the revised TFR for 2018 is 3.38 children per Saudi woman. See Figure 3

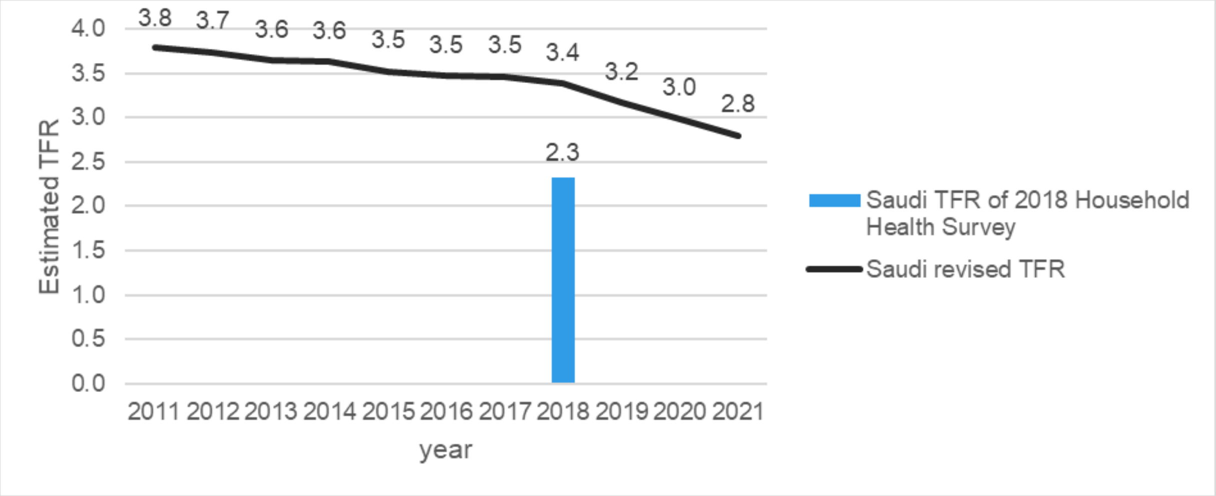

Figure 3: Comparison between the Saudi revised TFR verses the 2018 Saudi TFR from the Household Health Survey.

The estimated Saudi life expectancy at birth E(0) and the estimated Saudi death total over time are shown in Figure 4. The graph illustrates their relationship, with the death total rising in 2020–2021 and reflecting in a falling of E(0). Based on the estimated E(0), it is clear that improvements are being made throughout time, but that COVID-19 caused a fall in 2020–21 as the E(0) recovers post the pandemic.

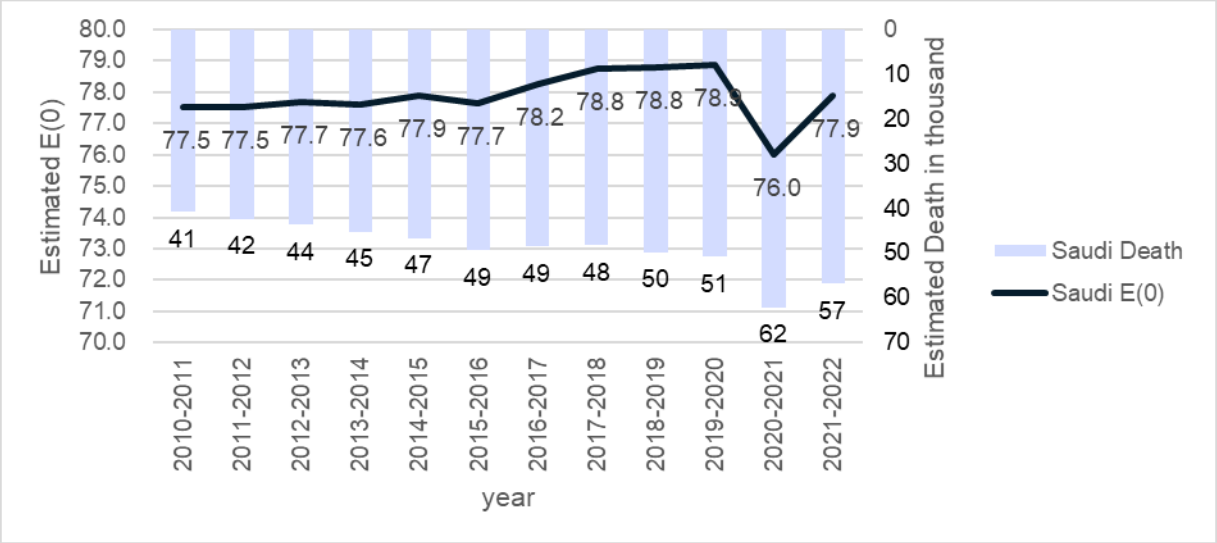

Figure 4: Comparison between the estimated Saudi Life Expectancy at birth e(0) and estimated Saudi death