# Aggregation underestimates growth: Case study in Population Modelling for India

**Shreenivas Kunte and Om P. Damani**
Centre for Technology Alternatives for Rural Areas, IIT, Bombay
Address: CTARA Office, 2nd Floor, Electrical Annex Building, IIT Bombay, Mumbai, 400076
Email: svkunte@gmail.com, damani@cse.iitb.ac.in

Jensen's inequality implies that aggregation underestimates growth and overestimates decay compared to an uncoupled disaggregated approach. As a study in disaggregation, this paper extends previous work in India's top down population modelling. Top down estimates of vital parameters used in country level population modelling hide the wide dispersion and interplay between these parameters that sub-national population cohorts have. Differences in total fertility rates, mortality rates and the demographics between national level averages and sub-national actuals, for large and diverse countries such as India, lead to a significant gap in population growth forecasts at the two levels. Population projections from International agencies, that are based on top down estimates, should be read in the context of these limitations. Higher, country level, starting data points conceal population groupings that demand more policy attention. Disaggregated modelling, used in this paper provides a plug and play alternative to construct sub national population projections that are used to derive country level estimates. As an example, state level population data, based on India's latest, 2011 census, have been used to create a population projection for India. Disaggregated population modelling provides another insight: geographically focused policy planning for fertility programs.

**Keywords**: Disaggregated modelling, India, Population projection, Jensen's inequality

# 1 Introduction

Large and diverse population agglomerations like India constitute moving parts that can be significantly different. For example, Goa, a small prosperous state with a per capita GDP of INR 2,24,138 (USD 3502) has considerably different fertility and mortality rates compared to Bihar's per capita GDP of INR 36143 (USD 565) (Government of India, 2015).

## 1.1 Modelling challenges

System Dynamic modelling used in this paper offers a conceptually aligned, simpler alternative for modelling large populations (Sterman, 2000). But population measurement practices camouflage the nature of important variables necessary for population forecasting. Census taking in India has a rich 140 year plus history (Dyson et al, 2004), yet birth and death reporting mechanisms use the sample registration system which is but a sampling mechanism. In addition, the civil registration system has limitations as it does not register births for a portion of the population (Guilmoto et al, 2013).

Census systems upgrade may solve some of the issues. What may not go away anytime soon is the wide dispersion in fertility and mortality in India's different population cohorts. Variations in vital parameters that influence population trajectory demand a disaggregated approach to modelling.

## 1.2 Disaggregated Modelling

As shown in Figure 1, fundamentally, all growth models involve a simple "Stock" that grows and declines, as determined by a growth rate and decay rate respectively. For projecting the growth for a group, a modeler, can choose any level of aggregation for determining three crucial parameters: initial stock value, growth rate and decay rate (for population modeling this includes starting demographic structure, fertility rate and mortality rate). The level of aggregation is often dependent on data availability, tolerance for errors, the need for brevity and finally the research question associated with the modelling exercise. While an aggregated (top down) approach supports parsimony, such approach involves information loss. But a disaggregated approach can deliver better results (Lindquist, 1999).

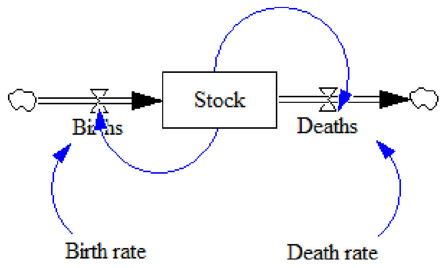


*Figure 1* A simple population model

For reference, aggregation biases, even for single equation models where the higher level aggregated variables are weighted averages of lower levels of the corresponding micro level variables, are not easy to estimate (Gupta, 1971). Since System Dynamic models involve non-linear equations, the complexity in the aggregation biases increases manifold.

## 1.3 Top down vs Disaggregated approach

Figure 2 is an extension of the simple model shown in Figure 1. The extension has one important change. Stocks for two distinct cohorts have been brought together to model a disaggregated projection. While remodeling Figure 2 scenario using a top down approach, the basic model shown in Figure 1 is reused as is but with one important change: Top down, weighted average values of growth (or decay) are used.

In contrast, in a disaggregated approach, as shown in Figure 2, there are two important changes. Although the basic model is the same, the model runs in tandem for two stocks. An auxiliary variable – “Total Stock” – sums up values of individual stock projections. The second important change in the disaggregated approach is that the critical variables for projecting growth in stocks are not (weighted) averaged top down. Instead, each model, uses a growth rate, decay rate and initial stock values (in case of population, the demography), that most closely represent the stock.

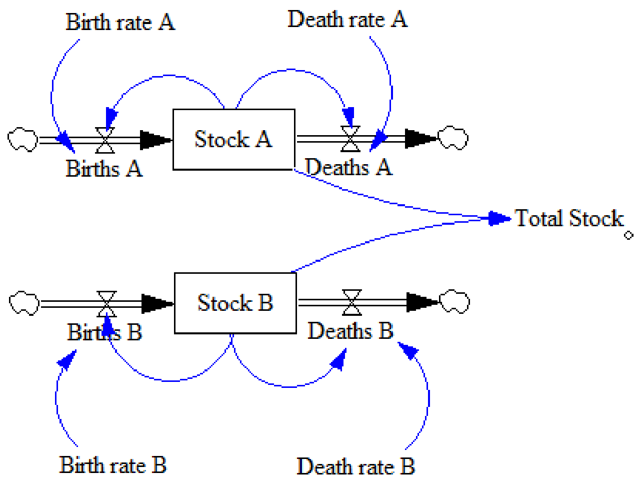


*Figure 2* Top down vs Disaggregated Approach

## 1.4 Comparing the two approaches – a generic case

Table 1 and Table 2 summarize the major differences between the top down and disaggregated approaches. In particular, Table 2 shows the effect of averaging critical parameters. Top down averaging of the net growth rate and the net decay rate (respectively) for stocks “A” and “B”, results in a net rate of 1.5%. Running the two stock projections in tandem (as shown in Figure 2) with critical parameters closely representing the underlying stock variables, results in a significantly different output.

| **Top Down Approach** | **Disaggregated approach** |
|---|---|
| No change in model | No change in basic model but two models run in tandem |
| Critical parameters averaged | Critical parameters most representative (not averaged) |

*Table 1* Top down vs Disaggregated approach differences

| | **Growth Rate %** | **Decay Rate %** | **Net Rate %** | **Starting Stock (Units)** |
|---|---|---|---|---|
| **People A** | 5 | 1 | 4 | 100 |
| **People B** | 4 | 2 | 2 | 80 |
| **Top Down** | 4.6 | 1.4 | 3.1 | 180 |

*Table 2* Specific values used in the model, increasing stock (net growth)

| | Growth Rate % | Decay Rate % | Net Rate % | Starting Stock (Units) |
|---|---|---|---|---|
| **People A** | 2 | 3 | -1 | 100 |
| **People B** | 2 | 4 | -2 | 80 |
| **Top Down** | 2.0 | 3.44 | -1.4 | 180 |

*Table 3* Specific values in the model, declining stock (net decay)

The contrast in model outputs between a top down and disaggregated approach is shown in Figure 3. The percentage gap between growth forecasts rises exponentially with time – from 11% in 50 years to about 45% in 100 years.

Table 3 and Figure 4 illustrate the difference between top down and disaggregated modelling when sub groups "A" and "B" (and the top down average) have declining net growth rates (net decay). It is notable that the trajectory of the "gap" curve in Figure 4 is identical to the trajectory of the "gap" curve in Figure 3.

This has an important implication – top down approaches to modelling underestimate growth and can lead to significantly different forecasts. As a corollary, it is important that forecasting exercises consider parameters (such as the fertility rate and mortality rate for population modelling) that closely represent the stock variables that are being forecasted.

## 1.5 Theoretical underpinning (Jensen's inequality)

Differences between top down and disaggregated approaches can be looked at in terms of Jensen's inequality (Weisstein, Mathworld). It states that the weighted mean of a convex function is greater than the function of the weighted mean:

$$f(w * x_1 + (1-w) * x_2) \leq \; w * f(x_1) + (1-w) * f(x_2)$$

Assume that stock *s* with initial value 1 is subdivided into stocks $s_1$, and $s_2$ with initial values *v* & *(1-v)* respectively. Let the corresponding growth rate of the 3 stocks be $g$, $g_1$, & $g_2$ respectively.

The value of any stock S with initial value I and growth rate r, after t time-steps is going to be

$$S(t) = I * (1+r)^t$$

Now consider the function:

$f(r) = (1+r)^t$

Since $f(r)$ is a convex function (it's second derivative $f''(r) \geq 0$), Jensen's inequality applies:

$$f(v * g_1 + (1-v) * g_2) \leq v * f(g_1) + (1-v) * f(g_2)$$

Note that $v * g_1 + (1-v) * g_2$ is nothing but top-down growth rate $g$ of stock. Therefore,

$$f(g) \leq v * f(g_1) + (1-v) * f(g_2)$$

Plugging in the value of function *f:*

$$1 * (1+g)^t \leq v * (1+g_1)^t + (1-v) * (1+ g_2)^t$$

Applying the definition of S(t), values of stocks s, s1 and s2 at time t are $s(t) = 1 * (1+g)^t$, $s1(t) = v * (1+g_1)^t, s2(t) = (1-v) * (1+ g_2)^t$, and therefore

$$\boxed{s(t) \leq s1(t) + s2(t)}$$

The Jensen's inequality simply implies that the top down growth rates will always underestimate the disaggregated growth rates. In case of decay, the inequality suggests that top-down approach will overestimate the decay compared to the disaggregated approach. The application of Jensen's inequality to system dynamics is not new (Montero, 2014). In macroeconomic forecasting, too it has been shown how aggregated models would fail to take into the dispersion in bottom up variables (Tevlin et al., 2003). But the relevance of Jensen's inequality principle provides an established framework to understand better top down models that involve growth, such as population forecasting.

Of course, it is important to realize that the simple derivation above assumes no coupling between stocks. Any real-life model, including population cohorts, will invariably involve coupling/feedback between stocks. In such cases, it may not be possible to make any clear-cut inferences about under or over-estimation due to aggregation. Considering this, the complexity introduced may not be worth the extra accuracy obtained by disaggregation (Randers, 1980). Further, as one drills down to lower level, micro-level inter-stock flows, which can be ignored at macro level, can no longer be overlooked. Depending on the purpose of the work, a suitable level of disaggregation needs to be decided. For example, for population projection, on moving from country to state level, one may drill down further and move down to district level. Most of the times this may not be desirable or feasible.

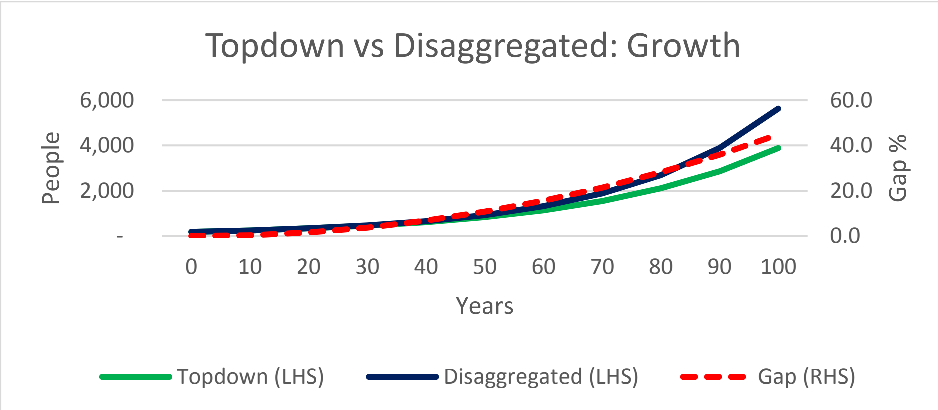


*Figure 3* Top down vs Disaggregated model output comparison – increasing stock (growth)

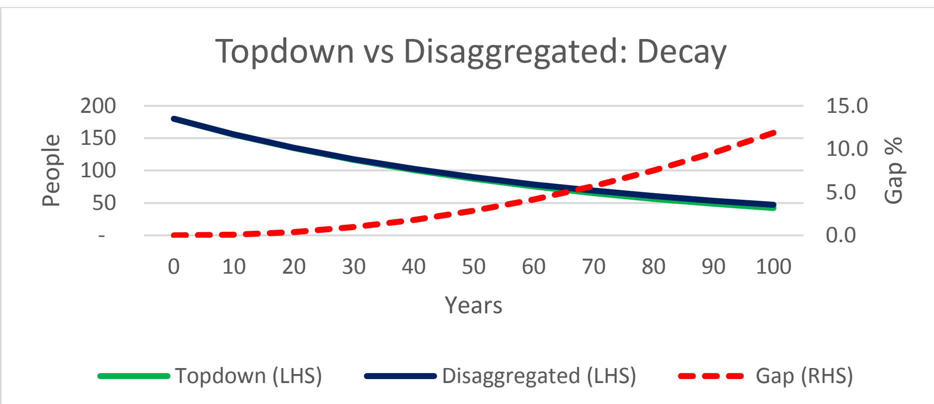


*Figure 4* Top down vs Disaggregated model output comparison – decreasing stock (decay)

# 2 Disaggregated India model & Data

## 2.1 Modelling India's Population

Population projections for India using a System Dynamics approach have been published before (Kunte et al., 2015). However, these projections were made using top down parameters. That is, the fertility rate, mortality rate and the demographic structure, took on values derived from top down averages. For instance, the fertility rate used in the population model, was an aggregated weighted average fertility rate for the entirety of India's national population. This paper improves on the earlier work by using a disaggregated approach to modelling India's population. Work in this paper also informs that SD based population models are easy to adapt at the micro level as well as corresponding macro, aggregated levels.

## 2.2 Model overview

The model in this paper, shown in Figure 5, provides inbuilt functionality for incorporating changes to key control variables as time progresses. Important control variables used in the model are: Fertility rate and mortality rate.

The twelve-stage model in Figure 5, tracks population progression of twelve age cohorts. In disaggregated modelling used in this paper, model illustrated in Figure 5 is used in tandem to project population trajectories of India's 35 states and union territories.

Nine distinct birth rates and six distinct mortality rates, for each of India's states & union territories have been applied to different classifications of age cohorts to determine population estimates.

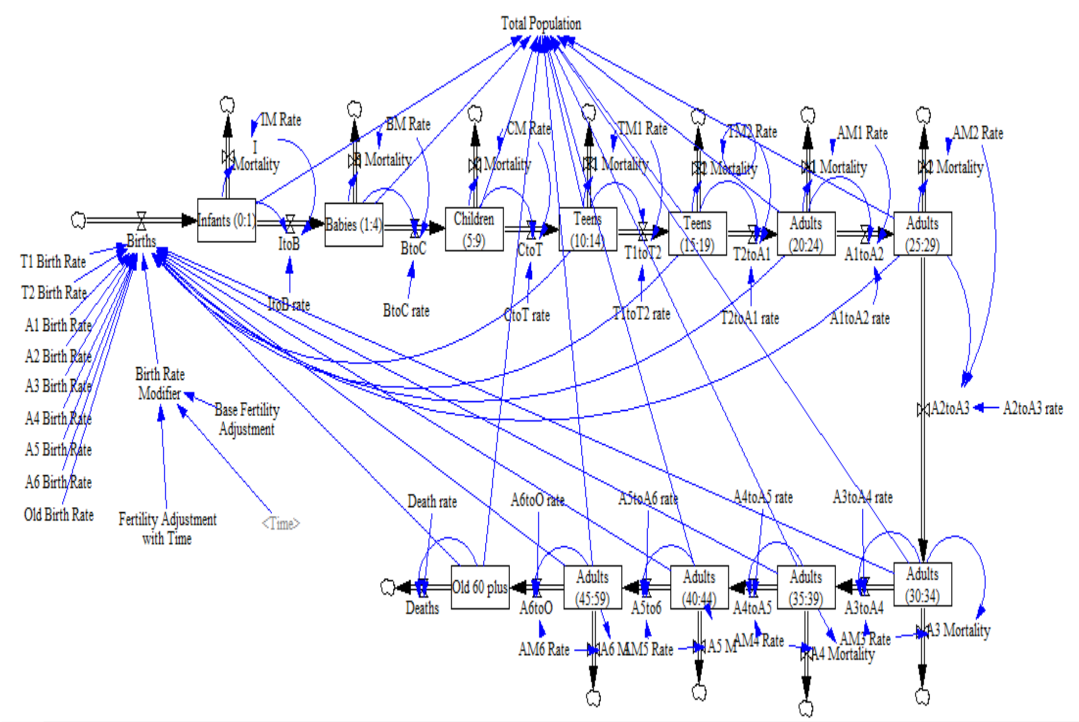


*Figure 5* 12 stage System Dynamics used in this paper

## 2.3 Data massaging and assumptions

Data massaging, approximations and assumptions made in the modelling exercise have been described below.

1) State level data takes precedence over the national level averages. Where available, state level data has been used for fertility rates, mortality rates and population demographic structure.

2) The model assumes a sex ratio of 1:1 for all age groups. To compensate for the fact that there are fewer women than men in India, the model calculates fertility per person in an age group and not fertility per woman. That is:

Age group fertility = (Total number of children born last year for this age group)/(Number of people in the age group)

3) The India Census data gives the crude death rate at an All India level (7 persons/ 1000). The Census data also gives a mortality weighting factor for each age bucket. Using the weighting factor and the total number of deaths, the model determines the mortality rate for each age bucket.

4) The model assumes same mortality rates for men and women. Infant mortality rates (41 per 1000) for male vs female (44 per 1000) are available via statement 45 of the census (India Census, 2012). But separate male vs female mortalities are not available for higher age groupings.

5) The “Births Last Year” data has two buckets for post menopause adults. The first bucket is for age groups between 45 and 59 years. The second bucket is for age groups above sixty years. Given relatively small absolute values (new born children assigned to this bucket) in this, the model combines the births in both the buckets to form a unified birth rate.

7) In order to incorporate the under reporting of babies born (but who have died) the model adjusts the fertility estimate upward by calculating an adjustment factor.

8) Unreported population is added back and spread equally across all other age buckets.

9) Crude death rates have been provided as ratios over designated ages ("< 1", "1-4", "0-4","5-14","15-59" and "60+"). Given that the 12 stage model has additional age cohorts, the model assumes same death rates for several age cohorts: children (5:9) and teens (10:14); all age groups between 15 and 44. Death rate for each state population cohort have been used where available.

10) Migration data is not available by age group. Overall, 7.6% of the population migrates within 4 years and 0.21% of the population have reported changing their residence every year. The model ignores migration effects.

11) In between 2001 and 2011, fertility decline in India has been rapid. At the national level total fertility rate came down by 0.5 points, from 3.16 to 2.66 (i.e. 15.8%). For India’s most populous states, UP and Bihar, fertility came down 17.7% and 6.6% respectively. The UN has

classified India as an "Intermediate Fertility" country (not "High" or "Low" – the two other categorizations). As shown in Figure 6, the UN's 2015 revision suggests a 6% reduction in fertility by 2025 (a medium fertility projection), (UN, 2015). The model in this paper provides a scenario analysis of population estimates based on future fertility changes. Base case assumption is for India to achieve a 50% reduction in totality fertility levels over the next decade (2011-2021), compared to the last decade (2001-2011).

12) The model assumes that after 2021, there would be no fertility decline. In reality, high fertility states will keep declining for a long time.

Note: Assumptions and approximations here are slightly different from the ones in (Kunte, 2015).

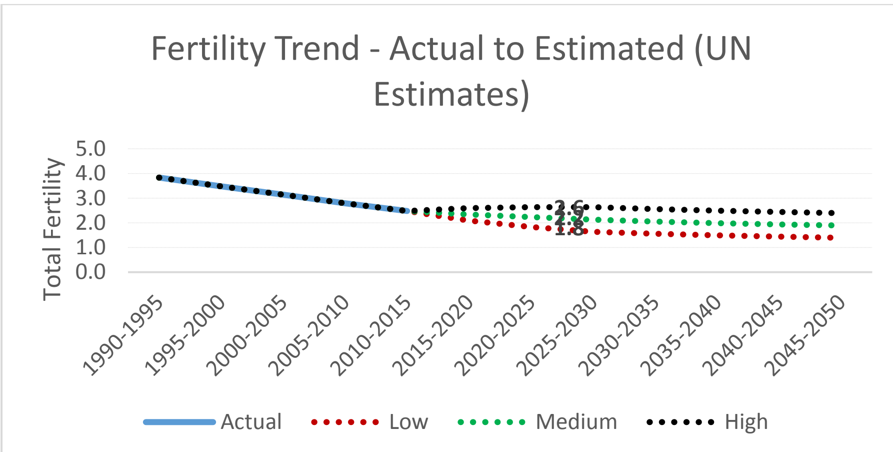


*Figure 6* Fertility actual and estimated trend for India. UN 2015 Revision

## 2.4 Input Data

Input data used for India's state level population projections has been sourced through India's 2011 census. Each state's demographic structure, along with fertility and mortality inputs have been used to develop the model. The data included in the paper can be referred to via India's 2011 census website (India Census, 2011).

# 3 Model output

## 3.1 Country Level Estimates

Country level estimates can be arrived at using two methods. As explained in Section 1.3, top down estimates are computed using national averages for fertility and mortality. Aggregated bottom up estimates are computed through adding up estimates for each the underlying groups.

Figure 7 and Figure 8 illustrate through charts the comparisons between top down and bottom up aggregated estimates of India's population.

The following observations can be made from the charts:

1) Bottom up aggregated projections made through summing up state level population projections are higher than the top down projections: over 20 years by 1%, over 50 years by 5% and over 100 years by 25%. Dyson et all (2004), highlight the use of sub national groupings to get better population estimates.
2) Bottom up projections are estimating India's population to grow by about 20% in 20 years and by about 38% in 50 years to 1.45 billion and 1.67 billion respectively (from 1.2 billion as of 2011 census)
3) India's four most populous and relatively backward states in terms of HDI, UP and Bihar, are estimated to grow by 78% (200 million to 355 million) and 106% (from 104 million to 214 million) respectively in 50 years (Sharma, 2015) (Sharma et al note that Bihar, Madhya Pradesh, Rajasthan and Uttar Pradesh selected as these states are amongst the most populous and are still considered to be lagging on development indicators)
4) If for some reason, fertility does not decline, over 50 years, India's disaggregated bottom up estimates rise by 47% instead of 42% using top down data (Table 5)
5) Table 4 shows the model output in comparison with other international agencies. It is notable that international agencies have a forecast of about 2.5% and 5.5% higher than the disaggregated model forecast over 15 years and 40 years respectively. While the exact fertility estimate used by the agencies is not known, given the tendency of top down models to underestimate growth, it follows that fertility estimate used by these agencies is higher compared to that provided by India's census.

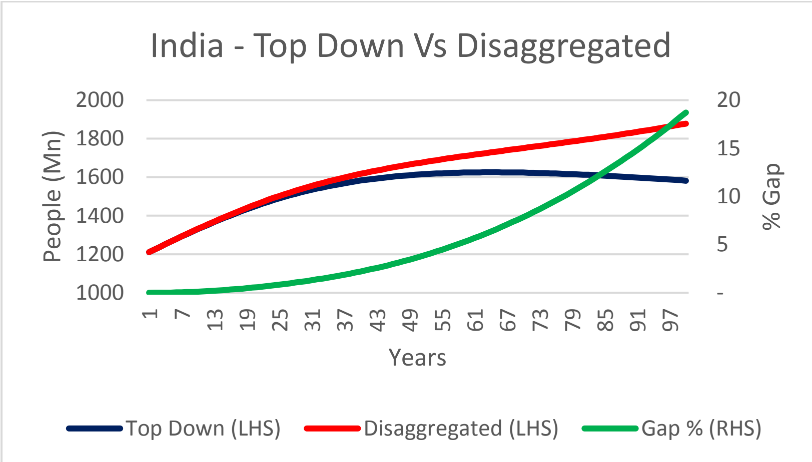


*Figure 7* Top down vs Disaggregated modelling differences

| **Million People / Year** | **2026** | **2050** |
|---|---|---|
| Model (Top Down) | 1390 | 1578 |
| Model (Disaggregated) | 1394 | 1611 |
| UN | 1430 | 1700 |
| WB | 1475 | 1705 |
| India Census | 1340 | |

*Table 4* Model output compared to estimates from other agencies

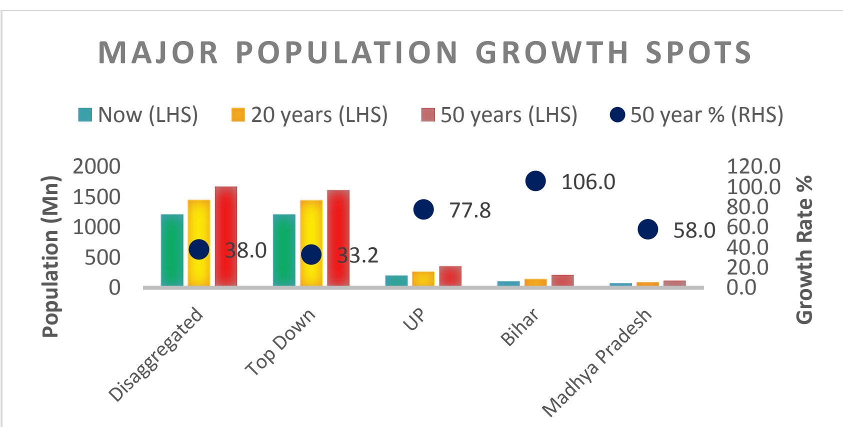


*Figure 8* Growth rates of India's most populous states vs disaggregated and top down growth rates

| Fertility Target | 20 Years | | 50 Years | | 100 Years | |
|---|---|---|---|---|---|---|
| | Top Down | Disagg. | Top Down | Disagg. | Top Down | Disagg. |
| No decrease | 21.3 | 22.0 | 41.9 | 47.0 | 55.5 | 84.2 |
| 10% (of 2001-2011) | 20.9 | 21.5 | 40.1 | 45.2 | 50.3 | 78.0 |
| 50% (of 2001-2011) | 19.2 | 19.8 | 33.2 | 38.0 | 30.7 | 55.1 |
| 100% (of 2001-2011) | 17.1 | 17.7 | 24.9 | 29.4 | 9.2 | 30.3 |

*Table 5* Population projection sensitivity in terms of growth rates for top down and disaggregated models for varying fertility targets

| Mortality Change | 20 Years | | 50 Years | | 100 Years | |
|---|---|---|---|---|---|---|
| | Top Down | Disagg. | Top Down | Disagg. | Top Down | Disagg. |
| 5% Down | 20.0 | 20.6 | 35.3 | 40.1 | 34.6 | 59.7 |
| 10% Down | 20.7 | 21.3 | 37.4 | 42.4 | 38.7 | 64.5 |

*Table 6* Population projection sensitivity in terms of growth rates for top down and disaggregated models for varying mortality targets

## 3.2 Sensitivity analysis

As illustrated in Figure 6, based on UN estimates, India's fertility rate is expected to decline by 6% in the next 10 years. Fertility in 12 states has gone below the replacement level of 2.1 (Guilmoto, 2015) but the backward states are still significant laggards. Fertility decline in the decade preceding the 2011 census has been used as a benchmark for targeted fertility decline over the next decade. Four differing sensitivity analysis scenarios used: 0% fertility decline compared to last decade (2001-2011); 10% fertility decline; 50% fertility decline and a 100% fertility decline.

Observations from the fertility sensitivity charts (Figure 9 and Figure 10):

1) At the country level, using aggregated state level data, Figure 9 suggests that a glide path of even a 50% of targeted fertility decline over the next decade is not enough for India's population curve to stabilize from an exponentially increasing function. But 50% decline in top down estimates is enough for India's population curve to stabilize. India's top down population estimates are more sensitive to fertility changes than population estimates based on aggregated data.
2) Population trajectories are quite insensitive to mortality changes (Figure 10)

Fertility sensitivity analysis information is more revealing at the state level.

1) Population trajectories of backward and most populous states are relatively insensitive to even a 50% (of targeted) fertility decline in a decade. In comparison, high HDI states (Dreze, 2012) are on a glide path for a secular decline for the base case sensitivity decline of 50% (of target) over the next ten years.
2) Even with a 50% fertility (of targeted) decline, population growth rates of the most populous states is significant.
3) On the other hand, with a base case of just 50% fertility decline over ten years, population trajectories of the high HDI states are peaking out (in case of Tamil Nadu, the state has already peaked out) in the next 50 years. High HDI states like Tamil Nadu could have a negative population growth over the longer term. The state's fertility rate has been on a decline for some time now. (Savitri, 1996).
4) It is to be noted that total fertility rates for Bihar, Rajasthan and Madhya Pradesh declined by 0.6% points in the period 2006 to 2011 (Kawadia et al., Aug 2014). While the states have been economically progressive, the ask rate for a more than 50% (of targeted) decline in fertility levels cannot be glossed over.

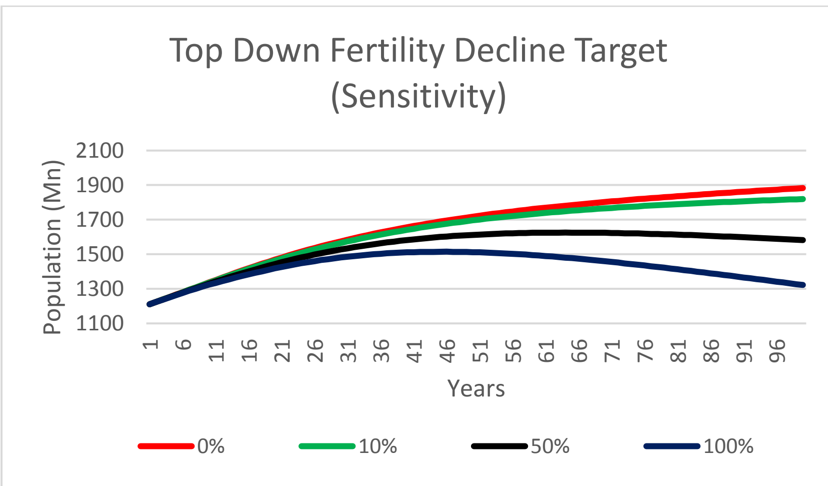


*Figure 9* Population projection sensitivity in terms of growth rates for top down model for varying fertility target decline relative to the decline in last decade

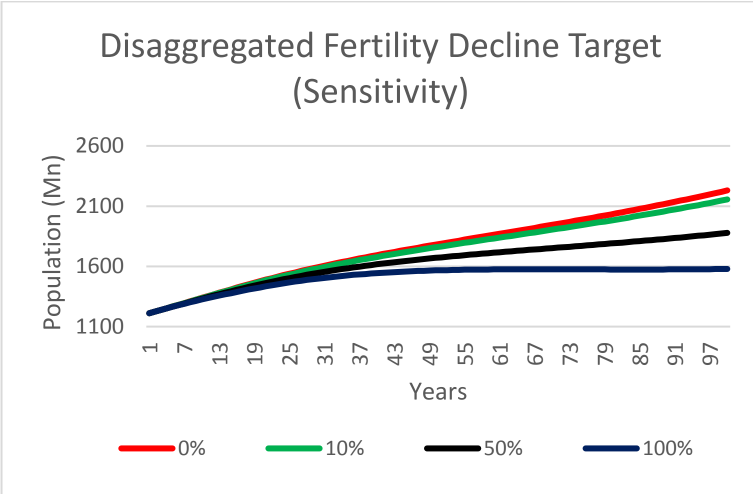


*Figure 10* Population projection sensitivity in terms of growth rates for disaggregated model for varying fertility target decline relative to the decline in last decade

## REFERENCES


Christophe Z Guilmoto, S Irudaya Rajan, Economic & Political Weekly. Fertility at the district level in India, Lessons from the 2011 census. June 2013.

Debraj Ray, Development Economics, Pages 60-61. 1998.

Estimates of Mortality Indicators, Census of India 2011, Registrar General of India, 2012.

Government of India, Press Information Bureau, Ministry of Statistics & Programme Implementation, http://pib.nic.in/newsite/PrintRelease.aspx?relid=123563, July, 2015.

Jean Drèze and Reetika Khera, Economic & Political Weekly. Regional patterns of human and child deprivation in India. September 2012.

John D Sterman, Systems thinking and modelling for a complex world, Business Dynamics, page 108. 2000.

Jorgen Randers, Elements of the System Dynamics Method, MIT Press, Page 35. 1980.

Juan Esteban Montero, Implication of the Jensen's Inequality for System Dynamic Simulations: Application to a Plug & Play Refinery Supply Chain Model, MS Thesis, MIT. February 2014.

Kanhya L. Gupta, International Economic Review, Vol 12/2. Aggregation bias in linear economic models. June 1971.

Kjersti-Gro Lindquist, Statistics Norway, The importance of disaggregation in Economic Modelling. June 1999.

Methods of population projection by sex and age, Manual III, UN, 1956

Michael Lipka, PEWResearch Center, Muslims & Islam: Key findings in the US and around the world. Feb 2017.

R Savitri, Economic & Political Weekly, Fertility Rate Decline in TamilNadu. July 1994.

Report of the Technical Group on Population Projections constituted by the National Commission on Population, May 2006

Robert Cassen, Leela Visaria, Page 1.Twenty First Century India, Tim Dyson, Page 1. 2004.

Sheena Sara Philips, Economic & Political Weekly. Shedding the BIMARU tag, Ganesh Kawadia. August 2014.

Shreenivas Kunte, Om Damani. Population projection for India – A System Dynamics approach. System Dynamics Conference. 2015.

Stacey Tevlin, Karl Whelan, Explaining the Investment Boom of the 1990s, Journal of Money, Credit and Banking. Feb 2003.

Tim Dyson, Robert Cassen, Leela Visaria. Twenty First Century India, Page 75. 2004.

United Nations, Department of Economic and Social Affairs, Population Division. 2015.

World Population Prospects: The 2015 Revision, Methodology of the United Nations Population. 2015.

Vinita Sharma, Economic & Political Weekly. Are BIMARU states still Bimaru? May 2015.

Weisstein, Eric W. "Jensen's Inequality." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/JensensInequality.html