Demographic Research Methods and the PyCCM library: Lecture Four

Charles Rahal and Jiani Yan

Leverhulme Centre for Demographic Science

A lecture delivered at the Banco de la Republica, October 2025

Motivation

- Convert 5-year age groups (0–4, 5–9, . . .) to single-year ages to estimate age-specific rates accurately.
- Enables life tables, annual CCM, and service planning by exact age.
- Reduces aggregation bias and aligns heterogeneous sources.
- Provides a coherent base for Colombia's departmental projections.

Problem setup

Fix a series (region, sex, year, variable). Unknown single-year counts

$$x = \left(x_a\right)_{a=a_{\min}}^{a_{\max}} \in \mathbb{R}^n_{\geq 0}.$$

Observed 5-year bands yield equalities

$$\sum_{a=\ell}^{u} x_a \;=\; b_{[\ell,u]}, \qquad u = \ell + 4.$$

Stack constraints as $Ax = b$ with $A \in \{0,1\}^{m \times n}$, typically $m < n$ (under-determined) $\Rightarrow$ add smoothness.

Design matrix: toy example

For ages 0–9 ($n = 10$) and bands $\{0\text{–}4,\, 5\text{–}9\}$ ($m = 2$):

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} B_{0\text{–}4} \\ B_{5\text{–}9} \end{bmatrix}.$$

Any admissible x must exactly recollapse to b by band.

Regularization via discrete curvature

Let D be the second-difference operator, $(Dx)_i = x_i - 2x_{i+1} + x_{i+2}$. Solve

$$\min_x \|Dx\|_2^2 + \lambda\|x\|_2^2 \quad \text{s.t.} \quad Ax = b,$$

with small $\lambda > 0$ for conditioning.

- Penalizes “wiggles” within bands; keeps totals exact.
- Clip to nonnegative, then re-project to the constraint plane if needed.

KKT system

$$\begin{bmatrix} 2(D^\top D + \lambda I) & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix}.$$

- Unique, stable solution when $\lambda > 0$ and $\mathrm{rank}(A) = m$.
- Post-solve: $x \leftarrow \max\{x, 0\}$ then re-project $Ax = b$.

Key properties

- **Conservation**: $Ax = b$ (preserve band totals exactly).
- **Smoothness**: minimizes curvature $\|Dx\|_2^2$ within bands.
- **Nonnegativity**: enforced with clipping and affine adjustment.
- **Robustness**: small ridge λ stabilizes near-collinear bands.

Splitting a single 5-year band

If only one band $[\ell, \ell+4]$ is known, allocate by weights $w_\ell, \ldots, w_{\ell+4}$ with $\sum w = 1$:

$$x_a = w_a \, B_{[\ell,\ell+4]}.$$

Choices:

- Uniform $w_a = \frac{1}{5}$ (uninformative baseline).
- Exposure-based $w_a \propto nL_a$ (life-table motivated).
- Empirical priors (historical single-year shapes by departamento).

Special treatment of 0–4

Use survivorship $\{l_0, \ldots, l_5\}$ and infant parameter a_0:

$$nL_0 = l_1 + a_0(l_0 - l_1), \qquad nL_a = \tfrac{1}{2}(l_a + l_{a+1}),\ a = 1, \ldots, 4.$$

Normalize $w_a \propto nL_a$ on $a = 0, \ldots, 4$ and allocate $B_{0\text{–}4}$ realistically.

- Optionally hold age 0 fixed when separate neonatal data exist.

Handling mixed fine and coarse constraints

Mixed inputs like

$$\{0,\ 0\text{–}4,\ 5\text{–}9,\ \dots\}$$

are supported:

- Keep point constraints (e.g. $x_0 = B_0$) as rows in A.
- Replace the residual of 0–4 with a 1–4 constraint.
- Optimization reconciles overlaps while preserving all equalities.

Open-ended tails adjacent to 5-year bands

- Standardize tails ($70+$, $80+$) into 5-year bands first:

$$70\text{–}74,\ 75\text{–}79,\ 80\text{–}84,\ 85\text{–}89,\ 90+.$$

- Allocate using geometric or exposure-based shares to preserve tail totals.
- Then unabridge each standardized band to single-year ages.

Numerical safeguards

- Choose $\lambda \in [10^{-8}, 10^{-4}]$ for stability without distortion.
- Guard against zero band totals in normalization steps.
- Enforce canonical label parsing and ordering before solving.
- Re-collapse after solving to verify exact conservation.

Algorithm per series

1. Parse labels; define single-year support; build A, b.
2. Add refined rows for 0–4 and tails if available.
3. Solve KKT; obtain x; clip and re-project.
4. Validate; write single-year ages to feed rates and projections.

Validation checks

- $\|Ax - b\|_\infty$ near machine precision.
- Fraction of ages with $x_a = 0$ (diagnostic only).
- $\|Dx\|_2$ smoothness by band; inspect outliers.
- Recollapse to published bands for spot checks.

Interpretation and tuning

- Larger $\lambda \Rightarrow$ smoother within-band profiles.
- Add inequality constraints (e.g. monotone nonincreasing at old ages) if needed.
- Prefer exposure-based weights for child ages; avoid spiking at age 0.
- Expect wider uncertainty where few bands inform narrow spans.

Micro example: adjacent bands

Given $B_{10\text{–}14}$, $B_{15\text{–}19}$:

$$\sum_{a=10}^{14} x_a = B_{10\text{–}14}, \qquad \sum_{a=15}^{19} x_a = B_{15\text{–}19}.$$

- Smooth allocation across 10–19 with exact conservation in each band.
- If $B_{15\text{–}19}$ is sparse, borrow shape from historical priors.

Summary of unabridging

- Constrained, curvature-penalized solve from 5-year to single years.
- Special handling: 0–4 exposure weights; standardized tails.
- Outputs feed life tables, rates, Leslie matrices, and policy indicators.

Why have population policies at all

- Rapid demographic change affects food prices, labour markets, housing, and political stability.
- For Colombia: uneven departmental growth, rapid urbanization, rural aging pockets, and migration shocks.
- Policy planning hinges on age-specific projections for schools, clinics, pensions, and transport.
- Rights-based framing since Cairo: focus on maternal health, poverty reduction, and individual choice.

Two narratives in tension

- **Implosion**: falling fertility $\Rightarrow$ smaller workforces, aging, fiscal stress; risk of skills shortages.
- **Explosion**: continued growth, unmet FP needs, environmental pressure; strain on urban services.
- Colombia sits between: low-fertility urban cores vs. higher-fertility rural/peripheral areas.
- Policy must be place-specific; aggregate averages can mislead.

Empirics that matter locally

- Asynchronous transition across departamentos; age structures differ markedly.
- Urbanization in Bogotá/Medellín/Cali vs. aging in rural Andean areas.
- Migrant inflows (e.g. Venezuelans) and internal displacement reshape demand quickly.
- Health burden shift to NCDs; need long-horizon care and prevention strategies.

Policy frameworks

- Balance resources, growth rates, human capital, and age structure dynamics.
- Use scenario analysis tied to explicit fertility, mortality, migration paths.
- Integrate NGOs and multilateral support with national strategies; align incentives.
- Emphasize monitoring: projections updated with new vital stats and census audits.

Intervention levers

- **Passive** adaptation (e.g. adjust pensions, school places) vs. **active** measures (e.g. FP expansion, rural service upgrades).
- Maternal/child health, adolescent SRH education, and labour formalization.
- Managed internal relocation: secondary cities, transport links, and housing supply.
- Data-driven targeting using single-year projections by sex and age.

Four instruments

1. Information & services: FP in primary care; counselling; adolescent outreach.
2. Laws & norms: maternity protections; anti-discrimination; age-friendly workplaces.
3. Taxes & transfers: childcare deductions; conditional cash transfers; rural health incentives.
4. Direct provision: childcare, clinics, rural sanitation, and eldercare capacity.

Developmentalists and family planners

- Jobs/productivity/education vs. rights-based FP; both interact through the demographic dividend.
- Evidence links schooling expansions to fertility decline with a lag.
- Optimal mix varies by place: urban cores need eldercare and labour matching; rural areas need access and girls' education.

Early wins and program design

- Mortality reductions through immunization, clean water, primary care, and road access.
- Vertical vs. integrated: short-run gains vs. system resilience; Colombia benefits from integrated primary care.
- Monitoring and evaluation: build feedback loops from admin data to adjust programs.

Relocation and internal migration

- Anticipate urban growth; invest in secondary cities to avoid megacity overconcentration.
- Improve rural service bundles (schools, clinics, broadband) to slow forced out-migration.
- Use single-year projections to time expansions in grades, beds, and transit capacity.

Lessons from aging contexts

- Pension pressure; design incentives for longer working lives where feasible.
- NCD prevention through primary care; shift from acute to chronic management.
- Managed migration can stabilize workforce size and skill mix.

Why projections not just forecasts

- **Projections**: "what if" under explicit demographic assumptions; used to test policies.
- **Forecasts**: "what will likely happen"; require probabilistic statements.
- We prioritize internal consistency and transparency over point prediction.
- Outputs can be wrapped in scenarios or uncertainty bands later.

From simple growth to age-structured reality

- Exponential growth assumes fixed crude rates and stable age shares—rarely true.
- Age structure emerges from fertility timing, survival, and migration by age.
- Annual single-year CCM tracks cohorts exactly and matches service ages.
- Recollapse to 5-year bands only for publication; not for computation.

Inputs required for annual CCM

- Base population n_t^F, n_t^M by single-year age (post-unabridging).
- Annual ASFR by single-year mother's age $[\alpha, \beta]$; normalized to TFR path.
- Annual survival probabilities S_x^F, S_x^M, including open age S_ω.
- Net migration vectors m_t^F, m_t^M by single-year age; scaled to geography.

Preparing single-year inputs

- **Unabridge** any 5-year inputs first; preserve totals and smooth within bands.
- **Harmonize tails** (70+, 80+) into 5-year bands, then unabridge to 90+.
- **Align ages and labels** across pop, deaths, births, migration.
- **Check SRB and age 0**: ensure daughters/sons split and first-year survival are coherent.

Constructing annual survival

- Compute $S_x = \exp(-m_x)$ or $S_x = 1 - q_x$ from life tables at single years.
- Open-age S_ω via exponential tail using e_ω or T-ratios.
- Smooth micro-spikes from small counts with moving averages where justified.
- Validate that $S_x \in (0, 1]$ and is monotone decreasing at extreme ages.

Constructing annual ASFR

- Estimate single-year ASFR directly or scale a baseline shape to a target TFR.
- Keep weights normalized (sum to TFR since width$=1$); avoid age gaps.
- Convert to daughters/sons using SRB and sex-specific S_0 (first-year survival).
- Document timing shifts (tempo) if cohort ages are changing.

Constructing migration vectors

- Build net migration by age and sex from flows; if only national, scale to departments via exposure.
- Shape by age: young adults, children (family moves), and retirees (return moves).
- Half-step convention aligns exposure with transitions; avoid double-counting.
- Check sign and magnitude against external benchmarks.

Why a Leslie matrix at all

- **One linear map for the whole year:** births into age 0 (first row) and cohort aging with survival (sub-diagonal) are captured in a single sparse operator $\mathbf{L}$.
- **Exact cohort tracking at single-year ages:** each age advances by exactly one year; no leakage between non-adjacent ages.
- **Transparent policy levers:** fertility only touches the first row; mortality only the sub-diagonal and the ω, ω cell; migration enters additively via a half-step.
- **Computationally light and auditable:** $\mathbf{L}$ is banded and sparse, so projections are fast and every nonzero has a demographic interpretation.
- **Consistent with annual, single-year inputs:** matches the unabridged pipeline (ASFR_x, S_x, SRB) and avoids artifacts from 5-year stepping.

Leslie idea in one picture

$$n_{t+1} = \mathbf{L}\, n_t \quad \text{(closed population)},$$
$$n_{t+1} = \mathbf{L}\Big(n_t + \tfrac{1}{2} m_t\Big) + \tfrac{1}{2} m_t \quad \text{(open population)}.$$

- First row: births into age 0 next year; zeros at pre-fecund ages.
- Sub-diagonal: cohorts age exactly one year with survival S_x.
- Lower-right diagonal: open-age retention S_ω.

Start with an empty female–female block

$$\mathbf{L}^{(FF)} \leftarrow \mathbf{0} \in \mathbb{R}^{(\omega+1)\times(\omega+1)}.$$

- We will fill $\mathbf{L}^{(FF)}$ in three passes: fertility row, survival sub-diagonal, open-age retention.
- Annual single-year steps ensure clean, sparse structure.

Introducing the Cohort Component Method (CCM)

- Population is divided into (synthetic) cohorts, divided into ages.
- In a closed population:
  1. Those aged $x+n$ in n years from now must be aged x years now **and** survive.
  2. Births are attributed to mothers.
- In an open population, migration matters. Closing it simplifies calculations.
- The CCM can be neatly expressed in the language of transition matrices.
- If using age groups of a year wide, we can project forward a year at a time.
- Lets begin with a female only closed cohort. The three things we need are:

  1. Starting population
  2. Survivorship data
  3. ASFRs.

Cohort Component Method: Survivor-ship

- If we have five-year age groups (the projection interval), we can express survivorship across consecutive ages as:

$$ {}_5N_x^F(t+5) = {}_5N_{x-5}^F(t) \times \frac{{}_5L_x(t)}{{}_5L_{x-5}(t)} \tag{1} $$

- **Class Quiz:** Who can interpret this equation in words? It's important!
- This formula is then applied to all age groups except for the youngest (births) and the oldest (open age interval).
- Lets talk about those two age groups next.

Cohort Component Method: Survivor-ship

- For the open interval, we need to combine survivors from two age groups.
  1. The number of women surviving *into* the open-ended interval.
  2. The number of surviving women *already* in the open-ended age group.
- Assuming the open-ended age group in the life-table began at an age 5 years older than that in the population:

$$ {}_{\infty}N_x^F(t+5) = ({}_5N_{x-5}^F(t) \times \frac{{}_5L_x(t)}{{}_5L_{x-5}(t)}) + ({}_{\infty}N_x^F(t) \times \frac{T_{x+5}(t)}{T_x(t)}) \tag{2} $$

- Or, we just multiply both ${}_5N_{x-5}^F(t)$ and ${}_{\infty}N_x^F(t)$ by $\frac{T_x(t)}{T_{x-5}(t)}$:

$$ {}_{\infty}N_x^F(t+5) = ({}_5N_{x-5}^F(t) +_{\infty} N_x^F(t)) \times \frac{T_x(t)}{T_{x-5}(t)} \tag{3} $$

Cohort Component Method: Births

- To compute the births produced during the projection interval we need information from the age-specific fertility rates: ${}_5F_x$.
- We compute the number of births produced during the projection interval by each age group of women.
- We then sum these over all age groups, and survive them to the beginning of the next age group.
- From all births – at least for the Female projections alone – we consider only females born using the SRB.
- Lets see an example of this on the following slide.

Cohort Component Method: Births

- The number of births to women of different ages is the ASFR (${}_5F_x$) multiplied by the person-years lived/exposure in the population age group in the projection interval.
- We calculate this as the average of the population at t and $t+5$:

$$B_x[t, t+5] = {}_5F_x \times 5 \times (\frac{{}_5N_x^F(t) +_5 N_x^F(t+5)}{2}) \tag{4}$$

- **Class Quiz:** What's this in words?

Cohort Component Method: Births (Cont.)

- And then we substitute in from equation (1) the value for ${}_nN_x^F(t+5)$:

$$B_x[t,t+5] = {}_5F_x \times 5 \times (\frac{{}_5N_x^F(t) + ({}_nN_{x-5}^F(t) \times \frac{{}_5L_x}{{}_5L_{x-5}})}{2}) \tag{5}$$

- Lets then re-arrange ever so slightly for clarity:

$$B_x[t,t+5] = \frac{5}{2} \times {}_5F_x \times ({}_5N_x^F(t) + ({}_5N_{x-5}^F(t) \times \frac{{}_5L_x}{{}_5L_{x-5}})) \tag{6}$$

- Note the subscript x on B_x, which can be thought of as children born to women starting the interval at age x.

Cohort Component Method: Births

- Then, we sum the births (B [t, t+t], eliminating the subscript) over the reproductive ages:

$$B[t, t+5] = \sum_{x=\alpha}^{\beta-5} \frac{5}{2} \times_5 F_x \times ({}_5N_x^F(t) + ({}_5N_{x-5}^F(t) \times \frac{{}_5L_x}{{}_5L_{x-5}})) \tag{7}$$

where α and β represent the upper and lower bound of the pro-creative span.

- Where have we seen something similar to this before?
- As we're only projecting a female population, we consider only daughters:

$$B^F[t+5] = \frac{1}{1+SRB} \times B[t, t+5] \tag{8}$$

Cohort Component Method: Births (Cont.)

- Then we survive the births through time t to $t+5$ since we are interested in the population aged 0 to 5, rather than number of births.
- To do this, we multiple through by by $\frac{{}_5L_0}{5\times l_0}$ as follows:

$$ {}_5N_0(t+1) = \frac{B[t,t+5] \times_5 L_0}{5 \times l_0} \tag{9} $$

- Purely in terms of female births, we can write:

$$ {}_5N_0^F(t+5) = B^F[t,t+5] \times \frac{{}_5L_0^F}{5 \times l_0^F} \tag{10} $$

- By plugging in the formulas from (7) and (8) we can cancel the 5s:

$$ {}_5N_0^F(t+5) = \frac{{}_5L_0^F}{2 \times l_0^F} \times \frac{1}{1+SRB} \times \sum_{x=\alpha}^{\beta-5} \times {}_5F_x \times ({}_5N_x^F(t) + ({}_5N_{x-5}^F(t) \times \frac{{}_5L_x^F}{{}_5L_{x-5}^F})) \tag{11} $$

Moving into Matrix Notation

- The CCM is best expressed in the language of transition matrices.
    - This *greatly* facilitates the use of computer applications.
- A transition matrix is a table with rows and columns showing the expected number of individuals who end up in a state with the label on the row.
- We need a way to project survival for people who are alive today.
- The best way to do this is using our existing lifetable data.
- We're going to take a very similar approach to before, where we utilize lifetables as a sequence of information which can be used to project.

Has everybody seen matrix notation before?

A matrix: set of numbers arranged in rows/columns to form a rectangular array:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} \tag{12}$$

What is the shape of matrix A? How about this column vector, B?

$$B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{13}$$

What shape is the product of C = A $\times$ B? What is that product?

$$C = AB = \begin{bmatrix} (a_{1,1} \times b_1) + (a_{1,2} \times b_2) + (a_{1,3} \times b_3) \\ (a_{2,1} \times b_1) + (a_{2,2} \times b_2) + (a_{2,3} \times b_3) \\ (a_{3,1} \times b_1) + (a_{3,2} \times b_2) + (a_{1,3} \times b_3) \end{bmatrix} \tag{14}$$

Survival Probabilities in Matrix Notation

- Note: diagonals are zero, because people don't stay at the same age!
- Below the critical first row, everything is a 'structural zero', **other** than lower subdiagonal.
- The probability of making it to N years later is stored in matrix element $\{j+1, j\}$:

$$A_{j+1,j} = \frac{{}_nL_{x+n}}{{}_nL_x} \tag{15}$$

- If there are ${}_nN_x$ people between age x and $x+n$ at time t, there will be the following number of people aged between x and $x+t$:

$$_nN_{x+n}(t+n) = A_{j+1,j}(t) \times N_x(t) \tag{16}$$

- Does this look familiar? Hint: equation (1)!

Introducing the Leslie Matrix

- We're building the (P.H.) Leslie (1945) matrix used for projecting the size and age distribution of a population through time:
  - By default, we project females only or 'females first' (female-dominant).
  - The matrix looks as follows, and we'll return to Births on the next slide:

$$\begin{bmatrix} {}_nN_0(t+n) \\ \dots \\ {}_nN_{x+n}(t+n) \\ \dots \\ \dots \end{bmatrix} = \begin{bmatrix} ? & ? & ? & ? & ? \\ A_{2,1} & 0 & 0 & 0 & 0 \\ 0 & A_{3,2} & 0 & 0 & 0 \\ 0 & 0 & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & ? \end{bmatrix} \begin{bmatrix} {}_nN_0(t) \\ \dots \\ {}_nN_{x+n}(t) \\ \dots \\ \dots \end{bmatrix} \tag{17}$$

- An $n \times n$ matrix multiplied by an $n \times 1$ vector: what shape is the product?
- Lower sub-diagonals are exact survivor-ship calculations we've seen before.

First Row of the Leslie Matrix: Births

- We again want to start from ASFRs for n years for females in the age group between x and $x + n$.
- We are again going to restrict births to just daughters.
- However, not all daughters survive until the end of the period, so we need to add a correction, just like before.
- Note: the upper left element depends on n and empirical knowledge about youngest ages of child bearing.
- Lets reinforce the fact that mothers are exposed to different ASFR during the interval as they age.
  - And they have to survive to give birth!

First Row of the Leslie Matrix: Births (Cont.)

- The net fertility contribution is: $\frac{1}{2}{}_nF_x + (\frac{1}{2} \times {}_nF_{x+n} \times \frac{{}_nL_{x+n}}{{}_nL_x})$
- And therefore, adjusting for survival, we get:

$$\underbrace{\frac{{}_nL_1}{2 \times l_0}}_{\text{Survivingship}} \quad \times \quad \underbrace{({}_nF_x +_n F_{x+n}\frac{{}_nL_{x+n}}{{}_nL_x})}_{\text{Number Born}} \tag{18}$$

- Adjusting for the SRB, we get the elements of the first row of the LM:

$$A_{1,j(x)} = \frac{1}{1+\text{SRB}} \times \underbrace{\frac{{}_nL_1}{2 \times l_0} \times ({}_nF_x +_n F_{x+n}\frac{{}_nL_{x+n}}{{}_nL_x})}_{\text{All surviving children}} \tag{19}$$

Projecting from the Leslie Matrix

- We now have our full Leslie Matrix:

$$\begin{bmatrix} {}_nN_0(t+n) \\ \dots \\ {}_nN_{x+n}(t+n) \\ \dots \\ {}_\infty N_x(t+n) \end{bmatrix} = \begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} & \dots & \dots \\ A_{2,1} & 0 & 0 & 0 & 0 \\ 0 & A_{3,2} & 0 & 0 & 0 \\ 0 & 0 & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & A_{j,j} \end{bmatrix} \begin{bmatrix} {}_nN_0(t) \\ \dots \\ {}_nN_{x+n}(t) \\ \dots \\ {}_\infty N_x(t) \end{bmatrix} \tag{20}$$

Which can be used to project forward:

$$N^F(t+n) = A^F \times N^F(t) \tag{21}$$

A verbose look at the LM

- What does the Leslie Matrix look like in practice?
- Where $k = \frac{1}{(1+SRB)} \times \frac{L_1}{2 \times l_0}$:

$$A = \begin{bmatrix} 0 & 0 & k(F_4 \frac{L_4}{L_3}) & k(F_4 + F_5(\frac{L_5}{L_4})) & \dots & 0 \\ \frac{L_2}{L_1} & 0 & 0 & 0 & \dots & 0 \\ 0 & \frac{L_3}{L_2} & 0 & 0 & \dots & 0 \\ 0 & 0 & \ddots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \frac{L_m}{L_{m-1}} & \frac{T_{m+1}}{T_m} \end{bmatrix} \tag{22}$$

- What's happening, for example, in the first rows?
- What key assumption are we making about m?
- What about the end of the procreative range?

Projecting from the Male Population

- Men's reproduction depends on women's fertility and their survival depends on a men survival-only type of matrix.
  - That is, our lower diagonal and closing terms remain identical, other than switching to the years-lived to men.
- Survival refers to men (for children) and to women (for mothers).
- Men's births can be expressed as:

$$A^{M}_{1,j(x)} = \frac{\text{SRB}}{(1+\text{SRB})} \times \frac{{}_{n}L^{M}_{1}}{2 \times l_0} \times ({}_{n}F_{x} +_{n} F_{x+n} \frac{{}_{n}L^{F}_{x+n}}{{}_{n}L^{F}_{x}}) \tag{23}$$

Projecting the Male Population

$$
\begin{bmatrix} {}_nN_0^M(t+n) \\ \cdots \\ {}_nN_{x+n}^M(t+n) \\ \cdots \\ \cdots \end{bmatrix} = \begin{bmatrix} A_{1,1}^M & A_{1,2}^M & 0 & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} {}_nN_0^F(t) \\ \cdots \\ {}_nN_{x+n}^F(t) \\ \cdots \\ \cdots \end{bmatrix} + \\
\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ A_{2,1}^M & 0 & 0 & \cdots & 0 \\ 0 & A_{3,2}^M & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & A_{m,m-1}^M & A_{m,m}^M \end{bmatrix} \begin{bmatrix} {}_nN_0^M(t) \\ \cdots \\ {}_nN_{x+n}^M(t) \\ \cdots \\ \cdots \end{bmatrix} \tag{24}
$$

- And we can write this as:

$$
N^M(t+n) = A^{MF} \times N^F(t) + A^{MM} \times N^M(t) \tag{25}
$$

Multi-year projections

- Having an initial population structure by age and sex ('base population') and a set of time-constant LM allows us to project the population forward in an *internally consistent* way.
- For women:

$$N^F(t+n) = (A^F)^t \times N^F(0) \tag{26}$$

- And for men:

$$N^M(t+n) = [(A^{MM})^t \times N^M(0)] + [(A^{MF})^t \times N^F(0)] \tag{27}$$

Migration

- Migration is usually added via a net migration vector (by age and sex).
- Alternatively: use separate vectors (as opposed to net) with immigrants and emigrants by age and sex.
- Preston et al. (2000) advise distributing half of the migrants at the beginning of the interval and the other half at the end.

Projecting Forward as a Forecast

- Is a projection like this a good forecast of the future?
  - Critically depends on assumptions on the input rates.
- However, we should expect demographers to be good at forecasting:
  - Population change occurs slowly compared to many other social processes.
  - Demographic processes have age profiles that are regular and change slowly.
  - We often have a long history of population statistics.
- Virtually all national-level forecasts are made by national statistical offices or international organizations (UN, Eurostat).
  - These forecasters are subject to bureaucratization and political pressure.

In reality, forecasts are bad!

thebmj covid-19 Research Education News & Views Campaigns Jobs

Feature » Data Briefing

Population projections: why they are often wrong

BMJ 2014 ; 349 doi: https://doi.org/10.1136/bmj.g5184 (Published 20 August 2014)
Cite this as: *BMJ* 2014;349:g5184

Article | Related content | Metrics | Responses

John Appleby, chief economist

Author affiliations

j.appleby@kingsfund.org.uk

Predicting the size of future populations is important for healthcare. Too bad our best guesses are often wrong, finds **John Appleby**

- However, demographic forecasts are typically quite poor!
- Charles (1938) saw the population of England and Wales declining from 41m to 18-29m by 2000 (Reality: 52m).
- Notestein (1945) predicted a total world population increasing from 2.5bn to 4bn by 2000 (Reality 6bn).

- Failed prediction of Baby Boom in the 1950s and 1960s/Lull in 1970s.
- Failed prediction of continuing trend toward lower mortality at older ages in (post-)industrialized countries.

Forecasting Mortality, Fertility and Migration

- An important issue in mortality forecasting is which indicator to forecast. Life expectancy is increasing, but age patterns are more complex. Attention is turning to the analysis of variations in lifespan.
- The UN assumes a very strong convergence in fertility due to the theory of Demographic Transition: i.e. fertility will increase where it is low and fall where it is high.

There are "no strong and consistent trends [in migration] that an serve as the backbone of credible projections for the future."

US National Research Council (2000), 'Beyond Six Billion'.

- Conventional projections may provide a small number of scenarios: 'High', 'Medium', and 'Low' Scenario-based approaches are very elementary, and more statistically sophisticated methods are becoming more common (e.g. fan charts).

Moving towards probablistic forecasts

- If a forecast consists of an individual number; almost certainly wrong.
- More statistical precision attainable with explicit confidence intervals.
- Three main approaches can be identified:
  1. Use of errors in past forecasts.
  2. Time series modeling (potentially multivariate, to incorporate GDP).
  3. Expert evaluation.
- Each of these approaches has its advocates; strengths and weaknesses.
- As yet there is no consensus on how best to estimate the uncertainty attached to forecasts.

Conclusions

- Projections are easy (if tedious); forecasting well is hard.
- If we look to the past to predict the future, how far back should we look?
  - Keyfitz and Caswell call this the 'experience base'.
  - Large body of econometrics looks at structural breaks.
- Errors come from uncertainty about trends in demographic processes of mortality, fertility, migration, but also from initial population enumeration.
  - Accurate quantification of birth rates and counting populations is challenging.
- Most probabilistic projections agree that uncertainty in the near future (next 20-25 years) is relatively small and most influenced by the uncertainty over future fertility. However, after about 25-30 years uncertainty becomes huge.
- This is because uncertainty over fertility runs into the second generation.
  - An uncertain number of mothers each has an uncertain number of births.

Multi-year updates

- Update ASFR paths (scale a fixed shape to target TFR) yearly, if policy scenarios require.
- Apply mortality improvement factors smoothly over time; recompute S_x as needed.
- Keep migration profiles current; shocks and policy changes matter most here.
- Recompute life tables if death counts or exposures change significantly.

Diagnostics and validation

- **Conservation**: compare recollapsed totals to official aggregates.
- **Plausibility**: compare age shapes to history and neighbours.
- **Stability**: track cohort lines across years (no unnatural waves).
- **Scenario differences**: attribute changes to assumptions (fertility vs. migration).

Sensitivity to assumptions

- Fertility: timing shifts vs. level; adolescent fertility changes have outsized effects on schooling.
- Mortality: improvements concentrated at older ages change pension ratios.
- Migration: transitory shocks vs. structural flows; age-shape matters for classrooms and labour.
- Document ranges; provide high/low/medium or probabilistic bands if feasible.

Conclusion

Unabridging to single years (inputs):

$$x = \big(x_a\big)_{a=a_{\min}}^{a_{\max}} \in \mathbb{R}^n_{\geq 0}, \qquad \sum_{a=\ell}^{u} x_a \;=\; b_{[\ell,u]}\ (u = \ell + 4), \quad Ax = b.$$

$$\min_x \ \|Dx\|_2^2 + \lambda\|x\|_2^2 \quad \text{s.t.} \quad Ax = b, \quad (\lambda > 0 \text{ small}).$$

Allocation within 0–4 when needed:

$$x_a = w_a\, B_{[\ell,\ell+4]}, \quad w_a \propto \begin{cases} nL_0 = l_1 + a_0(l_0 - l_1), \\ nL_a = \frac{1}{2}(l_a + l_{a+1}),\ a = 1, \ldots, 4. \end{cases}$$

Annual Leslie projection (single-year):

$$n_{t+1} \;=\; \mathbf{L}\left(n_t + \tfrac{1}{2}m_t\right) \;+\; \tfrac{1}{2}m_t,$$

$$\mathbf{L} : \begin{cases} \text{first row (births), zeros at pre-fecund ages,} \\ L_{x+1,x} = S_x \text{ (sub-diagonal survivals)},\ L_{\omega,\omega} = S_\omega. \end{cases}$$