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 $\{l_0, \dots, l_5\}$ and infant parameter a_0:

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

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

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

Handling mixed fine and coarse constraints

Mixed inputs like

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

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.

First row — fertility only

$$L_{0,x}^{(FF)} = \begin{cases} p_F\, S_0^F \cdot \mathsf{ASFR}_x^{\mathsf{annual}}, & \alpha \le x \le \beta, \\ 0, & \text{otherwise.} \end{cases}$$

- Early columns are **zero** because children and early teens are not fecund.
  - And, by extension, late columns are also zero.
- Optionally include a minor aging correction if ASFR varies steeply by age.

Add survival sub-diagonal

$$\forall x \in \{0, \ldots, \omega - 1\} : \quad L^{(FF)}_{x+1,x} \leftarrow S^F_x .$$

- This moves each cohort up exactly one age per year with survival.
- Off the sub-diagonal, entries remain zero (no skipping ages or leaks).

Add the lower-right diagonal

$$L^{(FF)}_{\omega,\omega} \leftarrow S^F_\omega .$$

- Intuition: those already in $\omega+$ at t remain with probability S_ω at $t+1$.
- Under annual steps, ω is typically 90; we use 90$+$ as open age.

Show the female–female matrix

$$\mathbf{L}^{(FF)} = \begin{bmatrix} \cdots & 0 & f_\alpha & \cdots & f_\beta & 0 & \cdots \\ \boxed{S_0^F} & 0 & \cdots & 0 & 0 & 0 & \cdots \\ 0 & \boxed{S_1^F} & 0 & \cdots & 0 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \\ \cdots & 0 & \cdots & \boxed{S_{\omega-1}^F} & \boxed{S_\omega^F} & & \end{bmatrix}$$

- Only the first row, sub-diagonal, and open-age diagonal are nonzero.

Add male births from mothers

$$L_{0,x}^{(MF)} = \begin{cases} p_M \, S_0^M \cdot \mathsf{ASFR}_x^{\mathsf{annual}}, & \alpha \le x \le \beta, \\ 0, & \text{otherwise.} \end{cases}$$

- Row in the male state that depends on the female state via mothers.
- Fathers' role enters indirectly through the exposure already in ASFR.

Add male survival

$$\forall x \in \{0, \ldots, \omega - 1\} : \quad L^{(MM)}_{x+1,x} \leftarrow S^M_x, \qquad L^{(MM)}_{\omega,\omega} \leftarrow S^M_\omega.$$

- Analogous to females: sub-diagonal aging and open-age retention.
- Sex differentials in survival matter at young and old ages.

Assemble the two-sex block matrix

$$\begin{bmatrix} n_{t+1}^{F} \\ n_{t+1}^{M} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{L}^{(FF)} & \mathbf{0} \\ \mathbf{L}^{(MF)} & \mathbf{L}^{(MM)} \end{bmatrix}}_{\mathbf{L}} \begin{bmatrix} n_{t}^{F} \\ n_{t}^{M} \end{bmatrix}.$$

- No $\mathbf{L}^{(FM)}$ term in a female-dominant model (births attributed to mothers).
- Can be extended to couple dynamics if needed; out of scope here.

Add migration with half-step convention

$$\begin{bmatrix} n_{t+1}^F \\ n_{t+1}^M \end{bmatrix} = \mathbf{L}\left(\begin{bmatrix} n_t^F \\ n_t^M \end{bmatrix} + \frac{1}{2}\begin{bmatrix} m_t^F \\ m_t^M \end{bmatrix} \right) + \frac{1}{2}\begin{bmatrix} m_t^F \\ m_t^M \end{bmatrix}.$$

- Keeps ages aligned with person-years of exposure; migrants accrue survival proportionally.
- Works cleanly with single-year ages and annual steps.

Matrix build recap — fertility only

$$\mathbf{L}_{\text{fert}} = \begin{bmatrix} 0 \cdots 0 & f_\alpha & \cdots & f_\beta & 0 \cdots 0 \\ & & \text{zeros elsewhere} & & \end{bmatrix}$$

- Daughters next year depend only on mothers this year at fecund ages.
- Age 0–12 columns are structurally zero in the first row.

Matrix build recap — add survival

$$\mathbf{L}_{\mathsf{fert+surv}} = \mathbf{L}_{\mathsf{fert}} + \begin{bmatrix} 0 & \cdots & \\ \boxed{S_0} & 0 & \cdots \\ 0 & \boxed{S_1} & \cdots \\ \vdots & \ddots & \ddots \end{bmatrix}$$

- Sub-diagonal moves everyone up one exact age under survival.
- No entries on the main diagonal for closed ages (no staying in place for a year).

Matrix build recap — add open-age retention

$$\mathbf{L}_{\text{full female}} = \mathbf{L}_{\text{fert+surv}} + \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boxed{S_\omega} \end{bmatrix}$$

- Ensures a realistic top age group under annual steps.
- With single-year ages, only the final cell on the diagonal is nonzero.

Tiny numeric illustration — fertility row (FF)

Annual step, single-year ages; fecund ages $x \in \{12, \dots, 49\}$.

Assume:

$$p_F = 0.488, \quad S_0^F = 0.990, \quad \mathsf{ASFR}_{12} = 0.020,\ \mathsf{ASFR}_{30} = 0.090,\ \mathsf{ASFR}_{49} = 0.010.$$

Then:

$$f_x = p_F\, S_0^F\, \mathsf{ASFR}_x \Rightarrow f_{12} = 0.0097, \quad f_{30} = 0.0435, \quad f_{49} = 0.0048.$$

Fertility row (nonzero only for maternal ages):

$$\begin{bmatrix} 0 & \cdots & \boxed{0.0097} & \cdots & \boxed{0.0435} & \cdots & \boxed{0.0048} & 0 & \cdots & 0 \end{bmatrix}$$

- Columns $0-11$ and $50+$ are zero — pre-/post-fecund ages.
- Each nonzero entry represents daughters born to mothers at that age who survive their first year.

Tiny numeric illustration

Add survival and open-age retention. Assume:

$$S_0^F = 0.990, \quad S_1^F = 0.995, \quad S_2^F = 0.996, \quad S_3^F = 0.997, \quad S_\omega^F = 0.940.$$

Sparse matrix with actual values (excerpt):

$$\mathbf{L}^{(FF)} = \begin{bmatrix} 0 & \cdots & 0.0097 & \cdots & 0.0435 & \cdots & 0.0048 & 0 & \cdots \\ \boxed{0.990} & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots \\ 0 & \boxed{0.995} & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots \\ 0 & 0 & \boxed{0.996} & \cdots & 0 & \cdots & 0 & 0 & \cdots \\ \vdots & & & \ddots & & & & & \vdots \\ 0 & \cdots & & & & \boxed{S_{\omega-1}^F} & \boxed{0.940} & & \end{bmatrix}$$

- Sub-diagonal $L_{a+1,a} = S_a^F$ moves cohorts up one age per year.
- Only the open-age cell (ω, ω) retains population.
- Combine with the previous fertility row for the full $\mathbf{L}^{(FF)}$.

Tiny numeric illustration (Cont.)

Add survival and open-age retention. Assume:

$$S_0^F = 0.990, \quad S_1^F = 0.995, \quad S_2^F = 0.996, \quad S_3^F = 0.997, \quad S_\omega^F = 0.940.$$

Sparse matrix with actual values (excerpt):

$$\mathbf{L}^{(FF)} = \begin{bmatrix} 0 & \cdots & 0.0097 & \cdots & 0.0435 & \cdots & 0.0048 & 0 & \cdots \\ \boxed{0.990} & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots \\ 0 & \boxed{0.995} & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots \\ 0 & 0 & \boxed{0.996} & \cdots & 0 & \cdots & 0 & 0 & \cdots \\ \vdots & & & \ddots & & & & & \vdots \\ 0 & \cdots & & & & \boxed{S_{\omega-1}^F} & \boxed{0.940} & & \end{bmatrix}$$

- Sub-diagonal $L_{a+1,a} = S_a^F$ moves cohorts up one age per year.
- Only the open-age cell (ω, ω) retains population.
- Combine with the previous fertility row for the full $\mathbf{L}^{(FF)}$.

Base-year assembly

- Choose base year t; ensure population, births, deaths, migration are **single-year** and aligned.
- If inputs are abridged: harmonize tails $\rightarrow$ unabridge $\rightarrow$ validate recollapse.
- Compute SRB and S_0 by sex from vital stats or life tables.
- Store metadata: sources, adjustments, smoothing parameters.

One-step run-through

$$n_{t+1} \;=\; \mathbf{L}\Big(n_t + \tfrac{1}{2}m_t\Big) \;+\; \tfrac{1}{2}m_t.$$

- Use $\mathbf{L}$ built from annual S_x and ASFR; confirm dimensionality.
- After the step, recollapse to 5-year bands for sanity checks only.
- Inspect age 0, teenage cohorts, and oldest-old for spikes.

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):

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

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