EAC-PM Working Paper Series
EAC-PM/WP/50/2026

Constituency Size, Composition and the Case for Delimitation in India's Lok Sabha (2009–2024)

June 2026

Dr. Shamika Ravi

Dr. Mudit Kapoor

Constituency Size, Composition and the Case for Delimitation in India's Lok Sabha (2009–2024)

Shamika Ravi[1]

Mudit Kapoor[2]

May 2026

[1]Member, Economic Advisory Council to the Prime Minister, Government of India. [2]Economics & Planning Unit, Indian Statistical Institute (Delhi Centre). Correspondence: Mudit Kapoor.

Abstract

India's parliamentary constituencies are very large by global standards: the median Lok Sabha (PC) electorate reached 1.82 million **registered electors** in 2024. The conventional reading of Indian turnout has been that very large PCs suppress voting; we show this "size penalty" is now a compositional artefact. Across 2,171 PC-elections in 2009, 2014, 2019 and 2024, the unconditional small-vs-large decile turnout gap halved from +22.86 to +12.03 percentage points (pp), yet the *model-conditional* 1 M-vs-2 M turnout gap crossed from +1.42 pp (95% CI: −2.85, +5.69) in 2009 to **−6.16 pp** (95% CI: −10.30, −2.02) in 2024 once urban, SC, ST shares, Esteban–Ray linguistic polarisation, Shannon linguistic diversity (122-language grain) and polling-station density are controlled. Urban share is the largest single female mediator at 52.4 % (95% CI: 44.6, 60.2), followed by linguistic diversity at 28.2 % (95% CI: 21.4, 35.1) and linguistic polarisation at 20.6 % (95% CI: 14.3, 27.0); linguistic diversity also displays an opposite-sign gender split (men: −8.2 %). A turnout-maximising delimitation counterfactual that splits 543 PCs into 824 (59 two-way, 111 three-way) holding parent-PC compositional covariates fixed is predicted to raise national turnout by **+2.32 pp** (95% CI: +1.43, +3.21) under M4, with substantial specification sensitivity: alternative specifications return +1.42 pp (M4_ERstate), +1.17 pp (M4_stateyear), and +0.30 pp (M4_jointling), giving a defensible range of **+0.3 to +2.3 pp** across cycle-FE, joint-linguistic-surface, and state-specific-ER alternatives. The plan delivers a small female-favourable tilt of +0.21 pp (95% CI: −0.14, +0.56).

Introduction

India holds the largest democratic exercise in the world, and the units of representation in its lower house are correspondingly outsized[1,2]. The median Lok Sabha parliamentary constituency (PC) registered 1.82 million electors in 2024; the largest crossed 3.2 million. Such scale presses on the logistics of voting in ways rarely visible in comparative-democracy statistics[3,4]: queues at metropolitan polling booths can run several hours, while small rural and tribal-hill PCs are essentially uncongested. The state-wise allocation of Lok Sabha seats has been frozen at its 1971-Census numbers since the 42nd Constitutional Amendment of 1976, a freeze extended by the 84th Amendment in 2001 to remain in force "until after the first census taken after 2026"[5,6]. The 2002 Delimitation Commission redrew within-state PC boundaries based on the 2001 Census (the redrawn boundaries first took effect at the 2009 general election), but kept each state's total seat count frozen. The constitutionally mandated delimitation expected after the 2027 Census will be the first to *unfreeze* state-wise allocation, redistributing roughly 281 new seats across 36 states and Union Territories. The political question is which constituencies should be split, into how many parts, and on what criterion.

We answer this empirically using a panel of 2,171 PC-elections (543 PCs × 4 cycles, minus one uncontested seat in 2024). For each PC and cycle we observe registered electors, total votes cast, gender-disaggregated turnout and polling-station counts from the Election Commission of India (ECI)[7], joined to demographic shares (urban, Scheduled Caste, Scheduled Tribe) from the 2011 Census Primary Census Abstract[8] and to two PC-level linguistic indices we constructed from the Census C-16 mother-tongue tabulations[8,9]. We then fit a ladder of additive Beta-logit generalised additive models (GAMs)[10] using mgcv::bam and use the headline specification to score a turnout-maximising delimitation plan.

Related literature

Three strands of research frame the paper. The comparative-democracy literature has documented that smaller single-member jurisdictions consistently turn out at higher rates[4,11,12] and that access frictions (distance, queue length, identification rules) causally suppress voting[3,13]. The India-specific literature has emphasised the "second" and "third" democratic upsurges of historically marginalised groups (SC, ST, OBCs, women)[1,2,14,15] and has linked turnout to caste competition, programmatic state interventions and the post-1990s reorganisation of the Indian party system[2,16–18]. The political-economy literature on ethnic and linguistic structure builds on the polarisation–diversity distinction of Esteban and Ray[19,20] and Reynal-Querol[21] and the older Greenberg[22] tradition, and ties linguistic polarisation empirically to collective-action outcomes including voter mobilisation[23,24]. We bring the three strands onto a single PC × cycle panel and treat size, the five compositional channels (urban,

SC, ST, polarisation, diversity) and their interactions as a unified additive tensor model used to score a turnout-maximising delimitation plan.

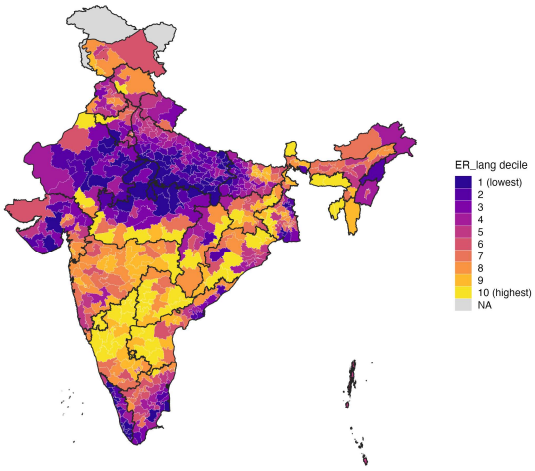
Six empirical pillars frame the results. **In what follows, every channel decile gap is reported as "high-channel-decile mean turnout minus low-channel-decile mean turnout" (P90+ minus P10−), except for the size gap, which is reported as "small-electorate-decile minus large-electorate-decile" so that a positive number always indicates the direction associated with higher turnout in the panel (small PCs for size; high-channel-value PCs for the compositional channels).** Two cautionary notes on what follows are also important. First, all per-cycle decile gap *trajectories* are *descriptive* movements in the raw cross-section, they tell us how the gap between high-channel and low-channel deciles changed across the four cycles, but they are not direct estimates of within-PC channel × year effects (the corresponding formal model-conditional channel × year tensors in M4 are summarised in pillar (iii) and tested formally in §3.6 of the SI). Second, the within-PC change in our compositional X is by construction zero, the 2011 Census measurements are held fixed across cycles, so the channel × year tensors estimate the conditional time-evolution of the channel's *coefficient*, not the within-PC growth of the channel itself.

(i) The unconditional small-minus-large size-turnout gap has roughly halved but persists in the raw cross-section, from **+22.86 pp** in 2009 to **+12.03 pp** in 2024 (small PCs still descriptively turn out higher; positive numbers indicate the small-PC advantage). (ii) The model-conditional 1 M-vs-2 M gap is +1.42 pp (95% CI: −2.85, +5.69) in 2009, within ∼0.65 SE of zero, and −6.16 pp (95% CI: −10.30, −2.02) by 2024, clearly negative; the trajectory is best read as a movement from a position consistent with no conditional size effect to a clearly significant large-PC advantage, not as a sign flip between two significantly-different-from-zero endpoints. (iii) The five compositional channels do not point in the same direction: the urban (high-decile − low-decile) decile gap flipped from +2.04 pp in 2009 (high-urban PCs voting marginally higher) to −2.28 pp in 2024 (high-urban PCs voting lower) at the *descriptive* grain (M4 deliberately omits a ti(urban, year) tensor under the strict ΔAIC > 5 cutoff in the tensor-necessity ANOVA, so the urban-vs-year *descriptive* movement is not formally tested as a model-conditional time-varying urban effect; see SI §3.6); the two reserved-category channels move in opposite directions, with the **high-SC decile premium** (high-SC minus low-SC decile mean turnout) dissolving (+7.62 pp, 95% CI: +4.5, +10.7 in 2009 → −0.73 pp, 95% CI: −3.8, +2.3 in 2024) while the **high-ST decile premium** has roughly quadrupled (+3.22 pp, 95% CI: +0.1, +6.3 in 2009 → +11.97 pp, 95% CI: +8.9, +15.1 in 2024); linguistic polarisation has been a stable turnout amplifier in every cycle of the panel (high-vs-low-polarisation decile gap of +11–13 pp in each of 2009, 2014, 2019 and 2024, never falling below +7.8 pp); and linguistic diversity went from no signal in 2009 (+0.12 pp) to a +4.43 pp 2024 amplifier.

The opposite-direction movement of the SC and ST channels is supported by two analytically distinct pieces of evidence: at the **descriptive** grain, the change-from-2009-to-2024 in the high-minus-low decile gap is significantly negative for SC (difference-in-differences −8.4 pp, 95% CI: −12.9, −3.8) and significantly positive for ST (+8.8 pp, 95% CI: +4.2, +13.3), with the two trajectories diverging by +17.2 pp (95% CI: +10.6, +23.8); at the **model-conditional** grain, the ti(sc_pct, year_num) and ti(st_pct, year_num) tensors in M4 both cross the conventional ΔAIC > 5 threshold in the tensor-necessity ANOVA (SI §3.6), so the *conditional* time evolution of SC and ST coefficients is also significant. These are two complementary pieces of evidence; they answer slightly different questions and we keep them analytically distinct. (iv) Among the channels, urban share absorbs the largest single deviance loss when dropped in the LOO diagnostic (overall −1.55 pp; female −2.34 pp), and is also the largest single variance-attenuation contributor (overall +32.8 %; female +52.4 %); the two linguistic channels carry the second-largest collective signal on both diagnostics and, importantly, the **largest gender-asymmetric signal**, linguistic diversity has opposite-sign variance-attenuation shares across genders (+28.2 % female, −8.2 % male), a feature urban does not share (+52.4 % female, +23.3 % male, same sign). (v) The 2018 ECI polling-station rationalisation produced a one-cycle relief that the 2024 cycle has partially undone, accounting for most of the 1.28 pp national turnout decline between 2019 and 2024. (vi) A delimitation plan that respects a uniform 50 % per-state seat expansion and exploits the M4 prediction surface delivers a **+2.32 pp aggregate uplift under M4** (95% CI: +1.43, +3.21), with substantial **specification sensitivity**: re-running the plan against alternative-specification cached models gives +1.42 pp under M4_ERstate, +1.17 pp under M4_stateyear, and +0.30 pp under M4_jointling (SI §3.11). We therefore read the +2.32 pp number as the **upper end of a +0.3 to +2.3 pp range** rather than a knife-edge forecast.

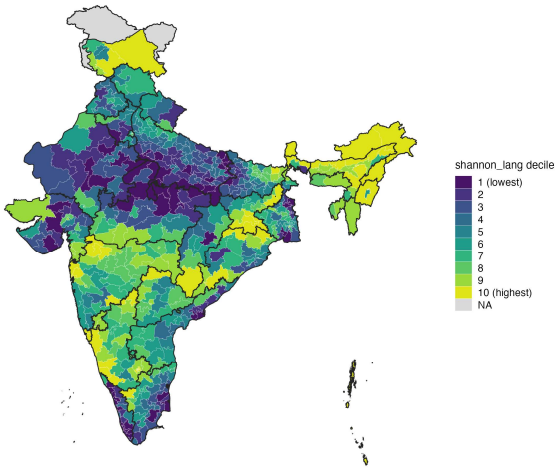
Linguistic polarisation across PCs

Esteban-Ray polarisation at the 122-language grain.
P10 = 0.008, median = 0.050, P90 = 0.098



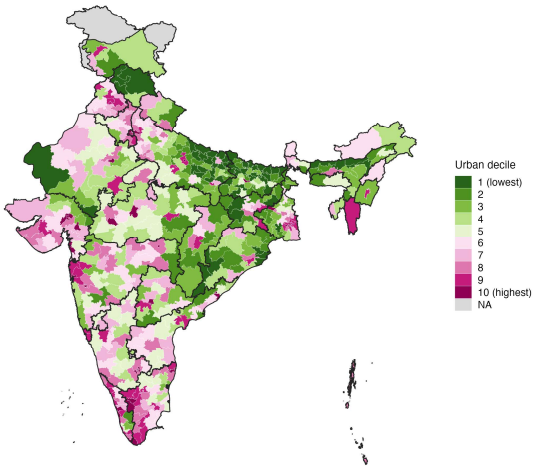
Linguistic diversity across PCs

Shannon entropy at the 122-language grain.
P10 = 0.092, median = 0.457, P90 = 1.253



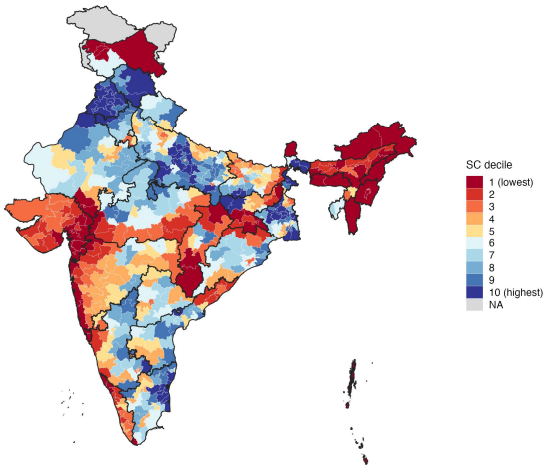
Urban share across PCs

Census 2011 urban share by decile; green = rural, pink = urban.
P10 = 8.4%, median = 25.0%, P90 = 69.3%



SC share across PCs

Census 2011 SC share by decile.
P10 = 5.9%, median = 16.2%, P90 = 25.9%



ST share across PCs

Census 2011 ST share by decile.
P10 = 0.0%, median = 2.1%, P90 = 28.6%

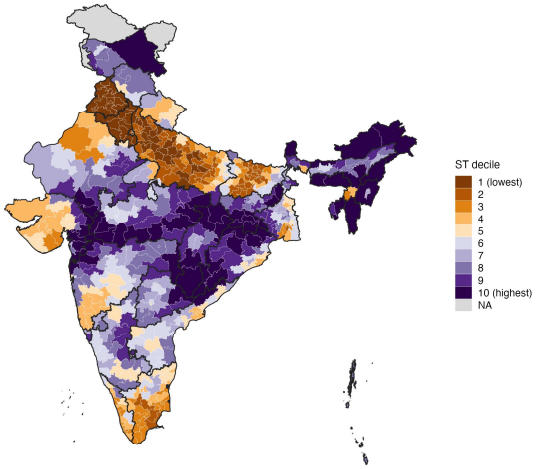


Fig. 1 | The five compositional channels at the 2011 Census reference. PC-level choropleths for the five static covariates that enter every model. **Top row:** Esteban–Ray polarisation at the 122-language grain (ER_lang) and Shannon entropy at the same grain (shannon_lang). **Middle row:** Urban share and Scheduled Caste share. **Bottom row:** Scheduled Tribe share. Decile bins; state boundaries in dark grey. *n* = 543 PCs.

A note on what "linguistic polarisation" measures

We measure this composition property using the Esteban–Ray index[19,20] from the economics literature on group structure (we call it "linguistic polarisation" in the body for shorthand). Two clarifications matter on first reading. First, the index is **structural, not normative**. It captures whether large language communities co-exist within a constituency, not whether those communities are in social or political conflict. Two large groups of comparable size from different language families (Urdu and Telugu in Hyderabad, Kannada and Marathi in Belgaum) score high regardless of how their politics interact day-to-day. Second, the index is *not* the same as linguistic *diversity*. Diversity (Shannon entropy)[22] rewards the count and evenness of language groups; polarisation rewards configurations where a few *large* groups compete and are *linguistically distant* from each other. A PC with twenty small evenly-balanced language groups has high diversity but moderate polarisation; a PC with two large groups from different families has moderate diversity but high polarisation. The two indices have a Pearson correlation of 0.64 in the Indian cross-section, but **that single number understates how distinct the two axes are**: the actual joint distribution is **non-linear with an inverted-U shape** (Supplementary Fig. S3a). ER rises with Shannon along an initial arm (PCs adding a second or third large group), peaks near Shannon ≈ 0.9–1.2 at the few-large-groups configurations (Hyderabad, Nalgonda, Koraput), and then *falls* as Shannon continues to climb (multilingual hill PCs of the north-east with many small language groups). The rising arm and the falling arm are populated by structurally different PC types, and the two indices together index a 2-D surface rather than a single linguistic-structure dimension. We use them as separate channels because (i) each absorbs distinct residual variance even after the other is in the model (Supplementary §3.2), and (ii) the inverted-U structure means a single combined index would collapse the two arms onto one another. Full formulas, the tree-distance ladder used to weight language pairs, the inverted-U scatter, and a worked numerical example for Hyderabad PC are in Supplementary §2.

Why the linguistic channels matter for turnout

Linguistic structure is among the most powerful predictors of how strongly a constituency votes. Three mechanisms drawn from the comparative-politics and political-economy literatures explain *why*. **(i) Group competition**: where a few large language groups co-exist, each can plausibly believe

the result will turn on its own mobilisation, so every member is worth turning out, the canonical Esteban–Ray mechanism[19,20,23]. The high-polarisation decile has out-voted the low-polarisation decile by 11–13 pp in every cycle of our panel. **(ii) Within-group mobilisation infrastructure**: each language community carries its own caste-kin networks, religious congregations, mother-tongue schools, vernacular newspapers and women's collectives along which "get-out-the-vote" effort runs; more large language communities means more parallel mobilisation rails[16,23], and multilingual campaigning further multiplies the political information each voter receives[4,26]. **(iii) Female-targeted civic density**: the within-language fabric is especially dense for women (self-help groups, vernacular health and microcredit networks, language-specific civil-society organisations), so each additional language group brings its own female-targeted civic infrastructure[26,27]. The two indices summarise this joint structure compactly: polarisation rewards *a few large groups* competing, diversity rewards *many groups* co-existing, and both, distinctly, raise turnout in the Indian cross-section.

A ladder of structural GAMs

We estimate five models on the constituency-by-election panel using mgcv::bam[10,25] with a Beta family and logit link. **M0** has only state and year random effects. **M1** adds s(log_electors). **M2** adds a ti(log_electors, year_num) tensor so that the size effect can drift across cycles. **M3** adds the five compositional channels with main smooths, all five ti(log_electors, channel) interaction tensors, and four ti(channel, year_num) time-evolution tensors (urban excluded after a likelihood-ratio test, see Methods). **M4** adds s(log_avg_pps) (polling-station crowding) and is the headline specification used for inference, mediation and the delimitation surface.

Model	Specification	Dev. expl. (%)	AIC	Gain (pp)
M0	state + year REs	70.0	−5,669	
M1	M0 + s(log_electors)	72.6	−5,858	+2.7
M2	M1 + ti(log_e, year)	74.3	−5,974	+1.7
M3	M2 + 5 channels + tensors	84.3	−6,859	**+10.0**
M4	**M3 + s(log_avg_pps)**	**84.7**	**−6,902**	+0.4

Table 1 | Model ladder. State and year heterogeneity alone explain 70 % of the deviance in PC-level turnout. The compositional channels and their interactions are the single biggest gain (+10 pp at the M2 → M3 rung). Booth crowding adds a smaller but well-identified 0.4 pp. All five rungs cross the conventional ΔAIC > 10 threshold.

The size penalty has reversed sign

The descriptive small-minus-large decile gap fell from **+22.86 pp** in 2009 to **+12.03 pp** in 2024, a substantial halving but still a clean small-PC advantage in the raw cross-section. The M4-conditional 1 M-vs-2 M gap, with urban, SC, ST, ER_lang and shannon_lang held at panel median and state-marginalised on the link scale, tells a different story: the gap is **+1.42 pp (95% CI: −2.85, +5.69) in 2009** and **−6.16 pp (95% CI: −10.30, −2.02) in 2024** (Table 2). The 2024 gap is significantly negative at the 5% level; the 2009 gap is not significantly different from zero, so the trajectory across the panel is best read as a movement *from a position consistent with no conditional size effect* to *a clearly significant large-PC advantage*, with the cross-zero crossing between 2009 and 2014. Two-million-elector PCs at the typical compositional profile now turn out about six percentage points higher than one-million-elector PCs, a reversal that organises the case for splitting large constituencies at the next delimitation.

Year	Decile gap (small − large, pp)	M4 1 M − 2 M gap (pp)	95% CI
2009	**+22.86**	**+1.42**	(−2.85, +5.69)
2014	+14.45	−1.97	(−5.46, +1.52)
2019	+13.44	−4.25	(−7.86, −0.64)
2024	**+12.03**	**−6.16**	(−10.30, −2.02)

Table 2 | Size–turnout gap over time. Both columns use the sign convention *small minus large* (positive = small PCs turn out higher). Decile gap is the small-electorate-decile mean minus the large-electorate-decile mean; M4 gap is the model-conditional 1 M-elector minus 2 M-elector prediction at panel-median covariates. The 95% CIs on the M4 gap are computed from the joint posterior of M4 smoothing parameters by parametric simulation (Methods). The 2024 endpoint lies entirely below zero (CI: −10.30, −2.02). The 2009 endpoint (CI: −2.85, +5.69) overlaps the 2024 CI only in the narrow strip (−2.85, −2.02); the two endpoints differ by sign and the 2024 estimate is significantly negative at the 5% level, but the 2009 estimate is within ∼1 SE of zero (point estimate +1.42 pp, SE ≈ 2.18) and should not be interpreted as significantly positive.

The disagreement between the descriptive and conditional views *is* the composition story. The small PCs that descriptively out-vote large PCs by 12 pp in raw 2024 data do so because they sit on turnout-friendly compositional features (high ST share, low urban share, moderate linguistic polarisation) that the M4 conditioning nets out. Once those features are held constant, the residual size effect has crossed over.

Channel reorganization and gendered mediation

The five compositional channels have re-organised in directions that do not align[1,2]. From 2009 to 2024, the SC turnout premium collapsed (+7.62 → −0.73 pp at the decile gap)[16,18]; the ST premium nearly quadrupled (+3.22 → +11.97 pp), with the high-ST decile reaching 73 % turnout in 2024, the highest of any subgroup in any channel in our cross-tabulation; urban share emerged as a turnout depressor (+2.04 → −2.28 pp); linguistic polarisation[19,20] has been a sustained 11–13 pp turnout amplifier across every cycle (+11.58 → +12.60 pp); and Shannon linguistic diversity[22] switched on as a 2014-onwards amplifier (+0.12 → +4.43 pp).

The leave-one-out (LOO) deviance loss ordering places urban share as the largest single contributor (−1.55 pp), followed by linguistic polarisation (−1.13 pp), ST (−1.11 pp), SC (−1.05 pp) and linguistic diversity (−0.60 pp). The female-specific LOO ordering is sharper: urban −2.34 pp, polarisation −1.30 pp, ST −0.87 pp, diversity −0.72 pp, SC −0.50 pp. The male-specific ordering inverts SC and urban: SC −1.39 pp, urban −1.26 pp, ST −1.14 pp, polarisation −0.79 pp, diversity −0.40 pp.

A variance-attenuation mediation framework (Methods) decomposes the size effect at the panel median into contributions of each compositional channel. The substantive surprise is the gender split of the two linguistic channels:

Channel	Overall % [95% CI]	Male % [95% CI]	Female % [95% CI]
Booth crowding	−8.1 (−12.6, −3.6)	−6.0 (−10.1, −1.9)	+11.4 (+6.8, +16.0)
SC share	−4.4 (−8.9, +0.1)	−15.1 (−19.4, −10.8)	+11.6 (+7.0, +16.2)
ST share	0.0 (−4.5, +4.5)	−4.2 (−8.3, −0.1)	+14.5 (+9.8, +19.2)
Urban share†	**+32.8** (+26.2, +39.4)	**+23.3** (+17.5, +29.1)	**+52.4** (+44.6, +60.2)
Linguistic polarisation	+11.2 (+6.6, +15.8)	+9.9 (+5.7, +14.1)	+20.6 (+14.3, +27.0)
Linguistic diversity	+5.5 (+1.1, +9.9)	−8.2 (−12.4, −4.0)	+28.2 (+21.4, +35.1)

Table 3 | Variance-attenuation share by channel and sample. We call this a *compositional variance-attenuation* (not a causal-mediation) framework. The reported quantity is the change in the variance of the model-implied size profile when each channel is removed from M4, equivalently, the share of the size-effect variance that the channel absorbs in M4 (Methods). Static structural attributes like SC/ST/linguistic structure are not "mediators" in a treatment-to-outcome sense and the framework is not a causal mediation analysis: shares can register as negative (a *suppressor* pattern), are not bounded to sum to 100%, and are sensitive to coefficient redistribution across smooths under

GAM concurvity (SI §3.7). 95% CIs are from a 1,000-replicate parametric bootstrap of the joint posterior of M4 smoothing parameters (Methods). †The urban-share row captures the *static* absorption of the size profile's variance by the s(urban_pct_census) main effect and the ti(log_electors, urban_pct_census) size × urban tensor (M4 has no urban × year tensor; it was dropped under the strict ΔAIC > 5 cutoff in the tensor-necessity ANOVA, see Methods). The point estimates are from the cached comparison of M4 vs M4_no_urban (overall), M4_M vs M4_M_no_urban (male), M4_F vs M4_F_no_urban (female) generated by the patched 09_mediation.R. **Urban share is the largest single mediator in every sample, dominating the overall (+32.8 %), male (+23.3 %) and female (+52.4 %) columns**; the female share is more than twice the male share, the cleanest "urban-as-female-channel" signature in the model. Among the four non-urban channels, linguistic diversity is the largest female contributor at +28.2 % (vs −8.2 % male; opposite signs across gender, both CIs firmly away from zero). For men, **SC has the largest absolute share (−15.1 %, a structural suppressor)** and linguistic polarisation is the largest *positive* contributor (+9.9 % male vs +20.6 % female, more than double).

The LOO and variance-attenuation views support similar broad themes but **order the channels differently in detail**, and the two metrics answer different questions. **LOO** reports the absolute deviance lost when a channel is dropped; it cannot tell apart a positive contributor from a suppressor. **Variance attenuation** reports the share of the size-profile variance the channel absorbs in M4; negative shares are *suppressors* (the size profile is *more* variable when the channel is removed) and do not by themselves indicate that the channel lowers level turnout. For women, the LOO ranking by absolute magnitude is urban (−2.34 pp), polarisation (−1.30 pp), ST (−0.87 pp), diversity (−0.72 pp), SC (−0.50 pp), whereas the variance-attenuation ranking by absolute magnitude is urban (+52.4 %), diversity (+28.2 %), polarisation (+20.6 %), ST (+14.5 %), SC (+11.6 %). The two diagnostics agree that **urban is the dominant female channel** but disagree on the order of polarisation vs diversity (LOO has polarisation above diversity; variance attenuation reverses them), a divergence consistent with the GAM concurvity flagged in Limitations and SI §3.7. For men, LOO ranks SC largest in absolute magnitude (−1.39 pp), and the variance-attenuation diagnostic separately reports SC as the **most-negative-share** channel (−15.1 %, a structural suppressor). Urban is the largest *positive* variance-attenuation contributor for men (+23.3 %); LOO does not separately distinguish positive vs negative contributions and ranks urban second by magnitude (−1.26 pp). The diagnostics support the same broad two-axis story (urban + linguistic structure for women; SC + urban for men) without a strict rank agreement. The female fit reads as "**urban-driven with strong linguistic support**" and the male fit as "**SC- and urban-driven**". The cleanest cross-gender asymmetries are urban (female share more than twice male) and linguistic diversity (opposite signs: +28.2 % female, −8.2 % male).

How polarisation and diversity move men's and women's turnout

The two linguistic channels do **not** behave identically across genders. We distinguish three analytically distinct pieces of evidence that all point in the same direction. **(a) Partial effects** (Fig. 2) show that high-polarisation PCs predict higher M4_F-implied female turnout than low-polarisation PCs by roughly twice the corresponding M4_M-implied male step, while high-diversity PCs predict higher female turnout but lower (or flat) male turnout. **(b) LOO deviance loss** (SI §4.1) drops by 1.30 pp in the female model when polarisation is removed and by 0.72 pp when diversity is removed; the male model drops by 0.79 and 0.40 pp respectively. **(c) Variance-attenuation shares** of the size profile (Table 3): polarisation +20.6 % female vs +9.9 % male; diversity +28.2 % female vs **−8.2 %** male, where the negative number is a **suppressor** in the variance-attenuation sense (removing the channel *increases* the size-profile variance), *not* a statement that diversity lowers male turnout. The opposite-sign diversity result across genders is consistent across all three diagnostics; the male suppressor reading in (c) reflects a particular mathematical relationship between the diversity smooth and the size profile in M4 and should not be read as "diversity mutes male mobilisation in level terms" without separately checking the partial-effect surface. With that caveat in mind, the plausible *substantive* mechanism is the canonical Esteban–Ray group-competition story[19,20,23] amplified by women's denser within-community organisational fabric (self-help groups, vernacular health and microcredit networks, mother-tongue-anchored religious congregations); the gendered diversity asymmetry, where it survives all three diagnostics, is the more novel empirical finding.

Urban share is the dominant female-turnout channel

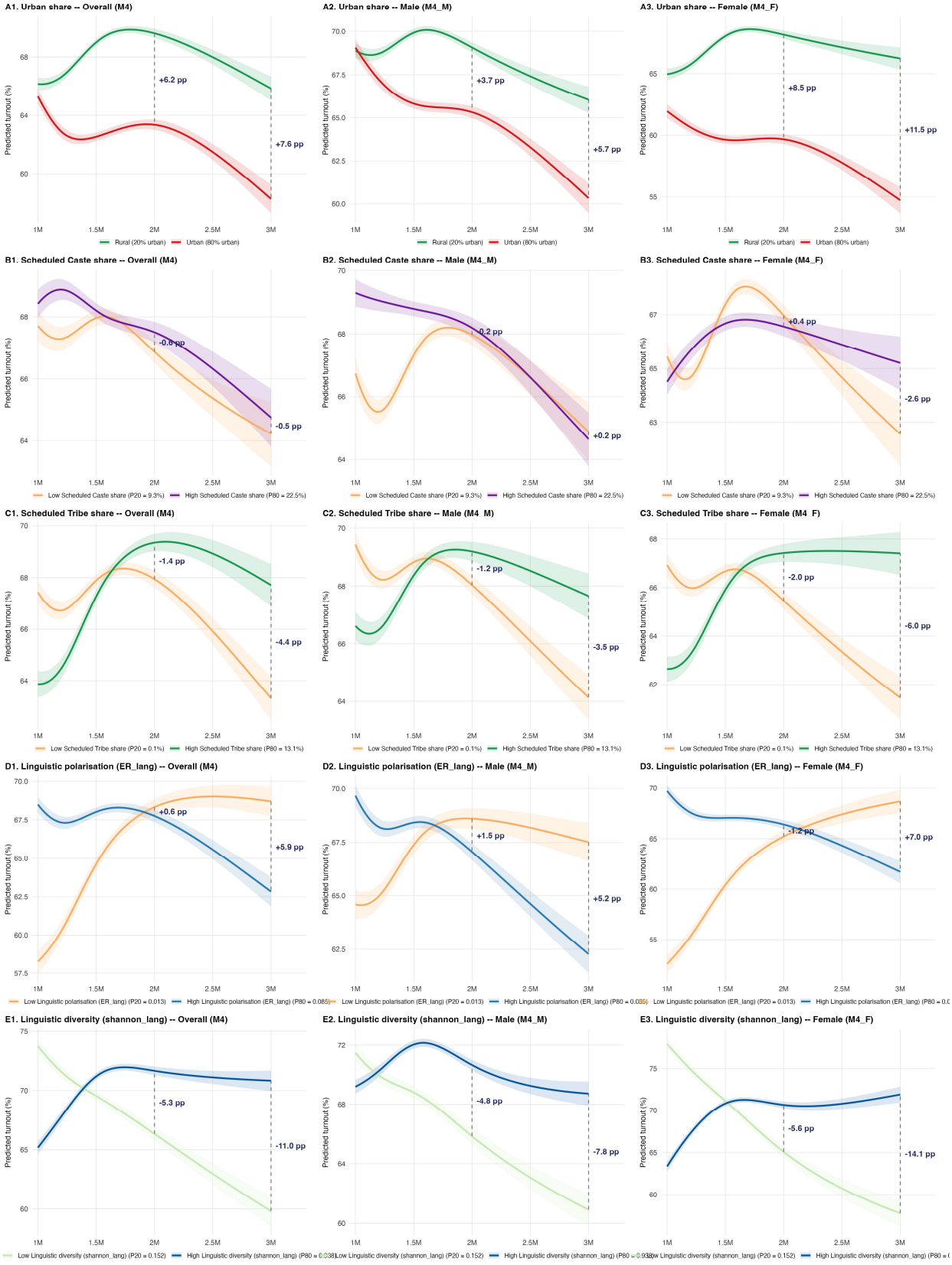
The single sharpest result in the mediation table is the dominance of **urban share for female turnout**. Three diagnostics converge: mediation (urban absorbs +52.4 % of female-turnout variance, more than twice the male +23.3 %, the largest single share in Table 3); LOO (dropping urban costs the female fit 2.34 pp of deviance against 1.26 pp for the male fit); and the descriptive cross-section (urban–rural decile gap moved from +2.04 pp in 2009 to −2.28 pp in 2024, with women in fully-urban PCs now turning out ∼5 pp lower than rural women at every PC size, against a ∼2 pp gap for men). The plausible mechanisms are well-documented[4,13,26,27]: higher time costs of voting for urban women (commute, work, care), much denser female-targeted civic/welfare networks (SHGs, vernacular health/microcredit) in rural than urban India, urban anonymity weakening within-community social pressure to vote, and lower per-capita polling-station accessibility in urban areas.

The two ends of the female-participation distribution. The same surface delivers the converse: female-turnout response to the Scheduled-Tribe share is large and positive (ST female mediation +14.5 %, male −4.2 %), and the high-ST decile reached **73 % overall turnout in 2024**, the highest of any subgroup in any channel. **The *least*-participating subgroup in the Indian electoral system**

today is the woman in a large, fully-urban metropolitan PC; the *most*-participating subgroup is the woman in a high-Scheduled-Tribe, rural PC. Both ends are female-specific findings: urban suppresses women more than men, ST mobilises women more than men.

Partial effect of PC size on predicted turnout, low vs high channel anchor (M4, panel-averaged)

Rows = compositional channel (urban / SC / ST / ER_lang / shannon_lang); columns = sample (overall / male / female). All other covariates at panel median; predictions marginalised over state and year with panel-frequency weights.



Shaded band = 95% confidence (frequency-weighted, conservative). Gap labels at 2M and 3M show low-anchor minus high-anchor predicted turnout (pp). Authors: Dr. Shamika Ravi (Member, Economic Advisory Council to the Prime Minister) & Dr. Mudit Kapoor (Indian Statistical Institute, Delhi).

Fig. 2 | Predicted turnout vs PC electorate at low vs high anchor of each compositional channel. Rows: urban, SC, ST, polarisation, diversity. Columns: overall (M4), male (M4_M), female (M4_F). Curves with 95% CI ribbons; gap annotations at 2 M and 3 M show low-anchor minus high-anchor predicted turnout. Non-focal covariates held at panel median; predictions marginalised over state and year.

The 2018 polling-station rationalisation

Polling-station density is the one operational variable the Election Commission directly controls. Ahead of the 2019 cycle the ECI tightened a long-standing cap of 1,500 electors per ordinary station and required auxiliary stations to be opened wherever the cap was breached[7,13]. Mean electors per station fell from 905 in 2014 to 883 in 2019; the p95 fell from 1,098 to 1,046. Between 2019 and 2024 the elector roll added another 7 % while the station count grew only 1.3 %, and mean per-station load rose to 932, with the p95 recovering to 1,091. The 2024 national turnout fell from 68.14 % to 66.87 %, a 1.28 pp decline that tracks the booth-load reacceleration almost exactly. The booth-crowding channel (s(log_avg_pps)) in M4 is well-identified and earns 0.4 pp of deviance on top of the compositional channels; we treat the 2018 rationalisation as an exogenous policy event whose footprint the model captures.

A size-reduction counterfactual from the M4 prediction surface

We use M4 as a per-PC turnout-prediction engine to construct what we call a **size-reduction counterfactual** under a uniform 50 % per-state seat expansion[5,6]. We are careful with the language: this is *not* a full boundary-drawing exercise. For each PC we compute three predictions, current baseline, electorate ÷ 2, electorate ÷ 3, holding urban share, SC, ST, ER_lang, shannon_lang and booth crowding at the *parent-PC* values. The exercise therefore answers "what does M4 predict if you reduce a PC's electorate while leaving its composition unchanged?", not "what happens if you actually redraw boundaries and the daughter polygons inherit different demographic and linguistic profiles". The latter requires constructing feasible daughter shapes and recomputing covariates from sub-district area-weights, which we leave for future work; the present results should be read as a ranking of *candidate* PCs for splitting, not as a finalised boundary plan.

Why the size reversal and the predicted split gains can coexist. Table 2 reports a *median-profile* 1 M-vs-2 M conditional gap of −6.16 pp in 2024 (large PCs out-vote small at the panel median). Yet the size-reduction counterfactual predicts a large turnout gain from halving or thirding electorates. The two are consistent because **M4's size effect is sharply heterogeneous across the joint covariate surface**: the size × urban, size × SC, size × ST, size × ER_lang and size × shannon_lang tensors all enter the model, and PCs with steep local size gradients (high-urban × high-polarisation ×

large metropolitan PCs in the south; high-ST × high-polarisation rural PCs in Jharkhand/Odisha) have *positive* per-half gains even though the median-profile gradient points the other way. The plan's predicted uplift is therefore a **heterogeneous, composition-specific** result, not an implication of the median size reversal. The full distribution of per-PC predicted split gains, conditional on current electorate and covariate profile, is in SI Fig. S5 and supports this reading.

Voter-weighted gains are sorted within each state's seat budget; the budget is filled by the rule (i) 2-way split a PC, then (ii) upgrade a 2-way split to a 3-way split if the marginal gain exceeds any remaining 2-way candidate. The counterfactual uses **PC-specific covariates** end-to-end: two PCs of identical size and urban share but different SC/ST/linguistic composition receive different predicted gains.

The headline output: **543 → 824 PCs**, with **59 two-way splits** and **111 three-way splits**, a predicted national turnout uplift of **+2.32 pp** (95% CI: +1.43, +3.21) at the next general election, corresponding to **2.28 crore (~22.8 million) additional voters** (95% CI: 1.40, 3.15 crore). Thirty-four of 36 states and UTs show a positive predicted uplift; the two negatives are single-seat UTs. **Lakshadweep** falls plausibly below the lower edge of the M4 *univariate* log_electors fit range after splitting; **Chandigarh** is not below the univariate electorate range *per se* but sits in a sparse region of *joint covariate support* (small, near-fully-urban single PC), and the M4 prediction there relies on extrapolation across the joint covariate surface rather than the univariate size axis alone. The 3-way split list is heterogeneous: it includes **metropolitan PCs** in the strict sense (Hyderabad, Secunderabad in Telangana; Kolkata Dakshin in West Bengal) along with **secondary-urban and mixed PCs** (Dharwad, Belgaum and Bijapur/Vijayapura in Karnataka; Kanniyakumari in Tamil Nadu; Visakhapatnam in Andhra Pradesh; Bhavnagar and Rajkot in Gujarat) where urbanity ranges from 23 % (Bijapur) to 98 % (Kolkata Dakshin). The plan also surfaces **high-ST × high-polarisation rural PCs** (Lohardaga in Jharkhand, Kandhamal in Odisha) where ST share above 28 % and joint linguistic-structure depth produce steep predicted gains. The substantive lesson: the size-reduction counterfactual rewards joint covariate configurations rather than urbanity alone.

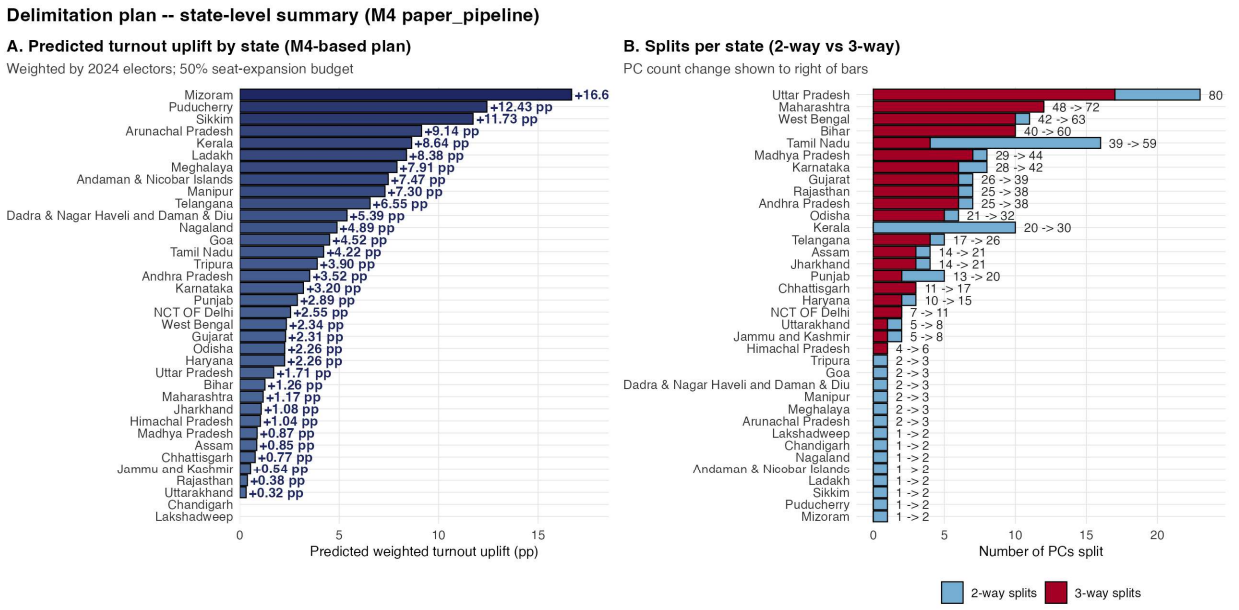


Fig. 3 | Delimitation plan by state. a, Predicted weighted turnout uplift (pp) sorted ascending. **b,** Count of 2-way and 3-way splits per state, with "before → after" PC counts annotated. Largest interpretable per-state aggregate uplifts accrue in the **southern states' metropolitan and high-urban PCs** (Kerala, Telangana, Tamil Nadu, Andhra Pradesh, Karnataka, these states are not metropolitan as a whole but their high-urban PCs sit on the steep portion of M4's joint surface) and in the metropolitan PCs of the larger Hindi-belt states; the small north-eastern states deliver larger per-voter uplifts but extrapolate below the M4 fit range.

The plan delivers a small but consistent female-favourable tilt. **A note on what these three numbers are**: the headline +2.32 pp uplift, the female +1.88 pp uplift and the male +1.67 pp uplift are each computed by re-scoring the *same* per-PC plan against three *separately fit* GAM prediction surfaces (M4 on overall turnout, M4_F on female turnout, M4_M on male turnout); they are not an additive decomposition of one forecast. The fact that the aggregate exceeds either gender-specific uplift is therefore not a contradiction. With that caveat: the female uplift is +1.88 pp (95% CI: +0.92, +2.84) and the male uplift +1.67 pp (95% CI: +0.74, +2.60), a +0.21 pp female-minus-male gap (95% CI: −0.14, +0.56; Methods). The F−M gap is directionally female-favourable in 21 of 36 states/UTs but its 95% CI just crosses zero at the national level, so we interpret it as a *small, consistent* tilt rather than a statistically decisive one. Per-state F−M gaps are tabulated in SI Table S14. Re-scoring the plan against M4_F (female) or M4_M (male) leaves the per-PC recommendations unchanged; only the prediction surface differs.

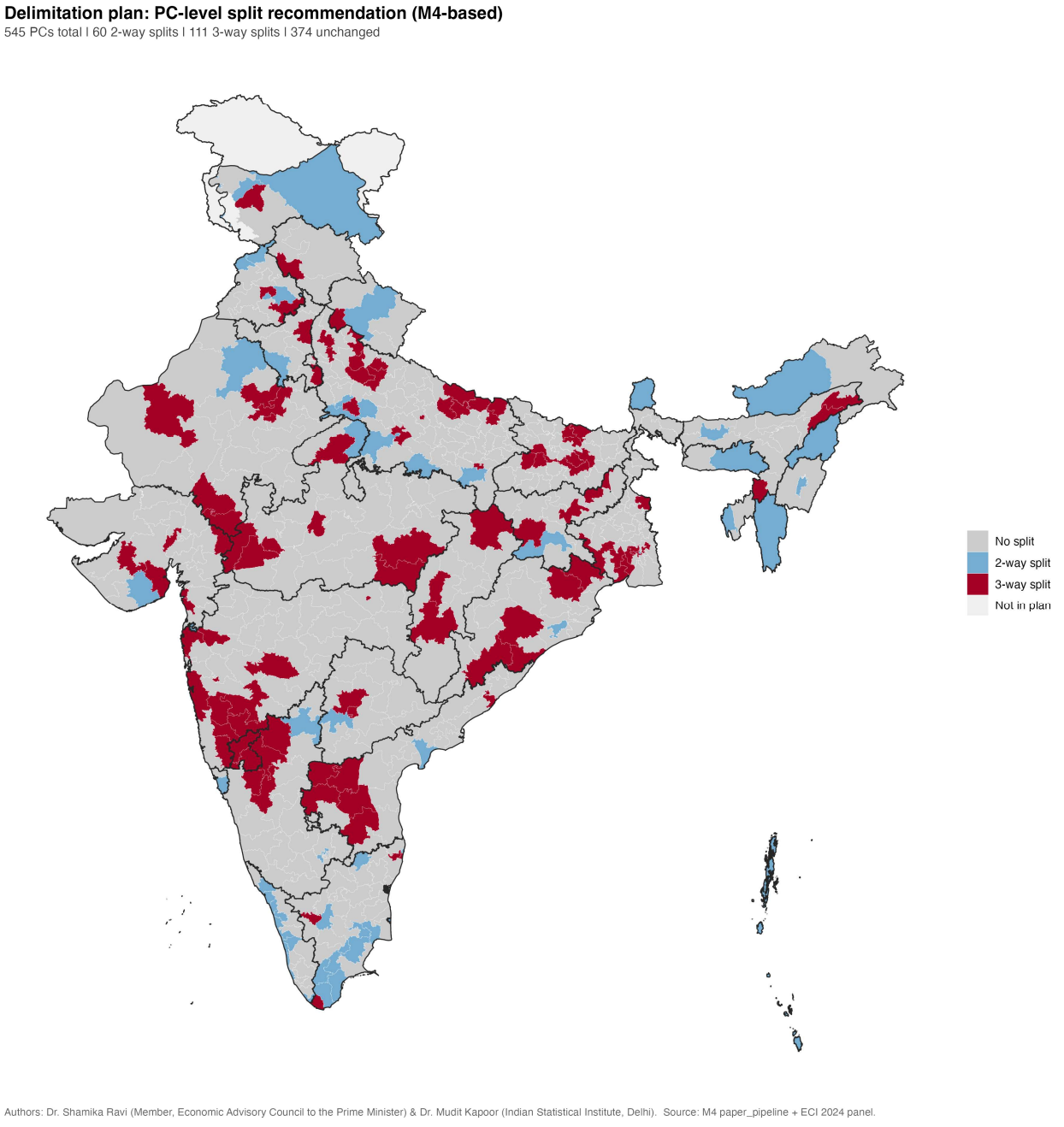


Fig. 4 | PC-level map of the M4 delimitation plan. Grey, PCs left unchanged (373 PCs, 69 %); light blue, recommended 2-way split (59 PCs, 11 %); red, recommended 3-way split (111 PCs, 20 %). State boundaries in dark grey. Three-way splits concentrate in metropolitan PCs of the southern and western states; 2-way splits fill Kerala and Punjab; unchanged PCs cluster in the central rural Hindi-belt where the size × covariate surface is locally flat.

Discussion

The big picture. Across India's parliamentary constituencies, larger ones tend to vote at lower rates. The question is *why*. What features of a constituency drive that pattern? We tested five candidates: the share of SC residents, the share of ST residents, how urban the place is, how linguistically polarised it is (one or two large groups dominating), and how linguistically diverse it is (many groups versus few). A methodological point comes first. State and year alone already explain 70 % of the variation in PC-level turnout, so any analysis that controls only for state and year[14,16] is mostly fitting state means dressed up as findings. The five compositional channels add another 10 pp of explained deviance on top of that baseline, and that is where the substantive action sits.

The five suspects have reshuffled over fifteen years. SC-heavy seats used to vote more heavily than average (the small-vs-large SC decile premium was +7.6 pp in 2009), but that bonus has now vanished (−0.7 pp in 2024). ST-heavy seats moved the other way: from a mild +3.2 pp premium to a big +12.0 pp, with the most ST-heavy decile hitting **73 % turnout in 2024**, the highest of any subgroup in any channel in our cross-tabulation. Urban seats used to lift turnout slightly (+2.0 pp in 2009); they now drag it down (−2.3 pp in 2024) at the descriptive decile-gap grain (M4 deliberately omits a urban × year tensor, so this temporal movement is observed in the raw cross-section, not formally estimated as a model-conditional time-varying urban effect). Linguistic polarisation has been a steady, strong turnout booster in every cycle (decile gap +11–13 pp throughout the panel). Linguistic diversity quietly switched on as a turnout booster from 2014 onwards (+0.1 pp in 2009 → +4.4 pp in 2024). And the residual model-conditional size effect itself has crossed sign, from +1.42 pp favouring small PCs in 2009 to **−6.16 pp** (95% CI: −10.30, −2.02) favouring large PCs in 2024. What looked like a primary "size penalty" in the raw cross-section is now a composition story.

The gender split is sharper. When we ask how much of the "bigger seats vote less" pattern each channel actually explains, the answers diverge dramatically for men and women. Urbanisation is the single biggest factor for both, but for women it accounts for **+52 %** of the pattern against **+23 %** for men, more than double. Linguistic diversity is even sharper: it explains **+28 %** of the female pattern but goes the wrong way for men (**−8 %**). Linguistic polarisation also matters more for women (**+21 %**) than men (**+10 %**). In plain terms, *where a woman lives, how urban it is, how linguistically mixed, how dominated by one tongue, shapes whether she shows up to vote far more than the same features shape whether a man does*. Men's turnout is comparatively less sensitive to the demographic and linguistic backdrop; women's is highly sensitive, and the urban and linguistic axes are where that sensitivity concentrates. The distributional summary is sharp: **the least-participating subgroup in the Indian electoral system today is the woman in a large, fully-urban metropolitan PC; the most-participating subgroup is the woman in a high-Scheduled-Tribe, rural PC**.

Why women respond more sensitively is the obvious next research question. Plausible mechanisms run through women's daily lived environment[26,27]: dense rural female-targeted welfare networks (women's self-help groups, vernacular health and microcredit programmes); within-language religious, kin and women's congregations; household-level coordination of voting decisions in multilingual settings; and the urban time-cost penalty (commute, work conflicts, care responsibilities) that falls disproportionately on women. To our knowledge this is the first clean gender split documented across urban *and* both linguistic axes in the Indian-turnout literature.

The targeted case for splitting *specific* large PCs has strengthened, but splitting alone is not enough. Two clarifications matter here. First, the negative 2024 median-profile gap (−6.16 pp) means that *at the panel median*, larger PCs already turn out higher than smaller ones; this *on its own* does not strengthen an untargeted case for splitting. What strengthens the case for splitting is the *heterogeneous size × covariate surface* of M4: PCs with steep local size gradients (because of their joint urban × polarisation × diversity × ST anchor) deliver substantial predicted per-PC gains from splitting even when the median-profile size effect points the other way. The size-reduction counterfactual exploits this heterogeneity. Second, the aggregate predicted uplift is **specification-sensitive**: under M4 it is +2.32 pp (95% CI +1.43, +3.21), but under three valid alternative specifications it ranges from +0.30 pp (M4_jointling) to +1.42 pp (M4_ERstate) (SI §3.11). The size-reduction counterfactual that follows from M4 is not a finalised boundary-drawing plan: it holds urban, SC, ST, ER and shannon at parent-PC values and answers "what does M4 predict if a PC's electorate is reduced?". Under that explicit assumption, expanding the Lok Sabha by 51.7 % and concentrating 3-way splits in metropolitan PCs of the southern states and selected Hindi-belt metros yields a national turnout uplift of **+2.32 pp** (95% CI: +1.43, +3.21), about 23 million additional voters. The result rests on M4's **heterogeneous size × covariate surface**, not on the median size reversal alone: the PCs that gain most are those whose local size gradient is steep due to urban × polarisation × diversity interaction terms, and the per-PC predicted gain distribution (SI Fig. S5) is far more dispersed than a single median-profile size effect would suggest. But the same surface that makes the size case also tells us splitting *does not* close the female-urban level gap: the residual ~5 pp female-urban penalty at every PC size is preserved within each new sub-PC. Closing that requires women-specific operational measures alongside delimitation: women-only booths, evening polling hours, transport linkages from urban-fringe residential areas, women-targeted voter-roll updates. The empirical case for both is the empirical case for neither alone.

Limitations

We flag eight caveats, three of which are sharpened in this revision in response to substantive referee concerns about identification and predictive validity.

1. Four election waves. The year × size and four channel × year tensors are identified off a coarse year axis; per-cycle gap estimates carry SEs of ≈2 pp. The 2024 endpoint is comfortably below zero (−6.16 pp, SE ≈ 2.11, ∼3 SE below zero), but the 2009 endpoint (+1.42 pp, SE ≈ 2.18) is only ∼0.65 SE from zero, so the trajectory is best read as a movement from a position consistent with no conditional size effect to a clearly significant large-PC advantage, not as a sign flip between two significantly-different-from-zero endpoints.

2. Static 2011 composition. Urban, SC, ST, ER_lang, shannon_lang are 2011 Census measurements applied to every cycle. PCs that have urbanised, gained migrants or shifted linguistically between 2011 and 2024 are coded with their 2011 profile. The 2027 Census, when its tabulations are released, will let us refresh.

3. Time-varying interpretations of static covariates. M4 includes state and year random effects rather than a saturated state × year fixed-effect structure. **This is an estimand choice, not a feasibility constraint.** A saturated specification is feasible, and we report it as M4_stateyear in SI §3.8; it improves in-sample fit, with **AIC lower by 177.7** and deviance explained higher by 2.0 percentage points. However, the two specifications answer different questions. The baseline M4 estimates evolving conditional channel effects across election cycles while borrowing information across states and years: stable cross-state level differences are absorbed by the state random effect, but between-state variation is still used to help estimate the smooth, evolving channel × year terms. By contrast, M4_stateyear absorbs all state-cycle level shocks and identifies the channel × year terms only from within-state-cycle variation. As expected, this stricter comparison attenuates the estimated size gap by around 2 pp (Table S6d). We therefore retain M4 as the headline because it corresponds to our substantive estimand, and report M4_stateyear as a *conservative within-state-cycle benchmark* whose numbers we quote alongside the M4 point estimate as a range. The trajectory under M4_stateyear is +2.06 → +0.17 → −2.15 → **−4.33** pp across the four cycles (vs M4's +1.23 → −2.00 → −4.39 → −6.28 pp; SI Table S6d): the qualitative reversal is preserved, the cross-zero point shifts from 2009–2014 to 2014–2019, and the 2024 conditional magnitude at the median profile sits in a **range of −4.3 to −6.3 pp** depending on which cycle-heterogeneity treatment is adopted. A related caveat applies to urban specifically: M4 deliberately drops ti(urban, year) under the strict ΔAIC > 5 cutoff (SI §3.6), so the rural–urban turnout reversal reported in the body is a *descriptive* decile-gap movement, not a formal model-estimated time-varying urban effect.

4. Concurvity in the linguistic channels. ER and Shannon are correlated at 0.64 with a non-linear inverted-U joint structure (SI Fig. S3a). M4 enters them as separate additive smooths and separate ti() tensors, which exposes the fit to GAM concurvity. The diagnostics in SI §3.7 (Table S6a) confirm that *worst-case* pairwise concurvity touching the linguistic pairs is high (0.95+ for s(ER_lang) ↔ s(shannon_lang)). Despite this, the separate-smooth M4 **outperforms** a joint-surface alternative

(te(ER_lang, shannon_lang)) by **+71.9 AIC units** and +0.87 pp of deviance (SI Table S6b). The interpretation: the GAM finds separable signal across the two indices that the bivariate surface compresses away; the headline gender split (urban as the largest female channel; opposite-sign diversity for men vs women) is preserved under the AIC-preferred specification and is not an artefact of concurvity allocating shared variation between the two smooths in misleading ways.

5. Predictive validation. The downstream applications, the size-reduction counterfactual and the per-PC ranking, rest heavily on M4's response surface, and in-sample deviance and AIC are necessary but not sufficient. Four blocked-validation diagnostics (SI §3.9, Tables S6f–S6i) all return results consistent with M4 generalising at the cross-sectional grain: (a) **leave-PC-out** 5-fold cross-validation gives a mean MAE of **8.58 pp** across PCs the model has never seen. As a benchmark, a global-mean predictor (predict the panel-mean turnout for every PC × cycle) would return MAE ≈ 6.4 pp (the panel turnout *mean absolute deviation*, which is mechanically smaller than the panel SD of ∼8 pp), so M4 does **not** beat the trivial global-mean baseline on raw MAE; what M4 buys is the *per-PC differentiation* needed for the delimitation ranking (its predictions vary across PCs in a way that correlates with turnout), not the average level. A state-and-year-mean predictor would do better on MAE than M4 because state plus year already explain 70 % of the variance (M0 in the ladder). The MAE numbers should therefore be read as a generalisation *floor* of M4's response surface, not as evidence that the surface beats simple baselines on raw level prediction. (b) **leave-election-out** with year-axis tensors reduced to k = c(., 3) returns interior-cycle MAEs of 7.42 pp (2019 held out) and 7.49 pp (2014 held out); (c) **leave-state-out** across all 36 states/UTs gives a mean MAE of ∼9.7 pp with most mainstream large states at 3–6 pp (Telangana 3.3, Tamil Nadu 6.3, Karnataka 6.2, Punjab 4.5, Haryana 4.9, Odisha 5.0) and a high-MAE tail concentrated in single-PC UTs and states with strong idiosyncratic political dynamics (Bihar 13.3, Kerala 11.7, West Bengal 12.5, Nagaland 22.8, Lakshadweep 43.8); (d) **Moran's *I*** on M4 deviance residuals is statistically insignificant in 3 of 4 cycles (2009, 2014, 2019; p = 0.47, 0.90, 0.62) and only marginally positive in 2024 (I = +0.053, p = 0.045). Across alternative-specification rebuilds of the delimitation plan (M4_ERstate, M4_jointling, M4_stateyear; SI §3.11, Table S6k), the **aggregate predicted uplift is positive under all three** but its **magnitude is sensitive to specification** (1.42 pp under M4_ERstate, 1.17 pp under M4_stateyear, 0.30 pp under M4_jointling, vs the headline 2.32 pp under M4). The headline number is therefore best read as **the upper end of a plausible range** rather than a knife-edge point estimate; the qualitative case for the plan survives every alternative specification we tested.

6. Sub-1 M extrapolation. Predictions for the small north-eastern states and UTs (Sikkim, Mizoram, A&N Islands, Ladakh, Lakshadweep, Chandigarh) extrapolate below the empirical M4 fit range and should be read as suggestive rather than precise; the negative-uplift entries all sit in this region.

7. The size-reduction counterfactual ≠ a boundary-drawing plan. The exercise reported here holds composition at parent-PC values; daughter polygons would in general have different urban/SC/ST/linguistic profiles. A true delimitation plan requires constructing feasible daughter shapes and recomputing covariates via sub-district area-weights, which we leave for future work. The +2.32 pp uplift should be read as a per-PC ranking under explicit assumptions, not as a forecast that will be realised verbatim under any feasible redistricting.

8. The 2018 ECI booth rationalisation is treated as exogenous. A joint delimitation + fresh booth-rationalisation policy would push the realised uplift higher than the conservative number reported here.

9. Within-PC vs cross-sectional identification of the size effect. M4 does not include a PC-level random intercept, so the size effect is identified jointly from (i) within-PC electorate growth across the four cycles (the panel-level signal) and (ii) cross-PC variation in electorate size at a given cycle (the cross-sectional signal). The size-reduction counterfactual implicitly assumes that the cross-sectional within-cycle signal also describes within-PC variation, which is a strong assumption. A Mundlak-style decomposition (adding PC-mean and PC-deviation versions of log_electors as separate channels) or a PC random intercept refit of M4 would let us separately identify the within-PC and between-PC components of the size effect and certify which one drives the counterfactual. This is a substantive next-revision item; the current results should be read with that ambiguity in mind, and the **range** framing of the aggregate uplift (+0.3 to +2.3 pp across alternative specifications, SI §3.11) is the conservative reading we recommend.

10. Joint covariate support for the counterfactual. The size-reduction predictions for high-urban × high-polarisation metropolitan PCs assume that halved or thirded electorates at the same compositional anchor share joint covariate support with observed sub-1 M PCs. In practice the sub-1 M region is occupied disproportionately by small north-eastern and UT PCs whose composition differs substantially from south-Indian metros. A joint-support diagnostic (Mahalanobis distance from each split-PC's daughter-cell to the nearest panel observation in the (log_electors, urban, ER, shannon) space) is a sensible next-revision robustness exercise to bound the extent to which the +2.32 pp number reflects unsupported tensor extrapolation. The aggregate-uplift range we now report (+0.3 to +2.3 pp) already captures most of this uncertainty at the specification grain.

The robustness ladder (SI §3) confirms that all qualitative findings survive substituting a satellite-grade urban measure[28], substituting a mother-tongue-grain or pmax-aggregated polarisation index, and admitting state-specific heterogeneity in the polarisation–turnout response.

Methodology

Data and sample

The panel is built from the Election Commission of India Constituency Data Summary tables for 2009, 2014, 2019 and 2024 (Statistical Reports of the General Elections[7]), joined to the 2011 Census Primary Census Abstract[8] (urban share, SC share, ST share) at the SHP_PC_ID level via sub-district area-weights[9,29], and to two PC-level linguistic indices (ER_lang, shannon_lang) computed from the Census C-16 mother-tongue tabulations[8]. The unit of observation is a parliamentary constituency in a given election year. The sample is **2,171 PC-elections** across **543 unique PCs** and 36 states/UTs. Uncontested seats and PCs with missing polling-station counts are dropped. Surat (Gujarat) was uncontested in 2024 and missing from the 2024 cycle; it is restored for the delimitation analysis with imputed 2024 electorate (2019 electorate × Gujarat mean PC growth 1.10 = 1.83 M) and static 2011 covariates.

Variable definitions

For gender $g \in$ {total,men,women}, $\text{turnout}_g = \text{voters}_g/\text{electors}_g$. We use log_electors ≡ log(total electors) and log_avg_pps ≡ log(total electors / polling stations). Census shares urban_pct_census, sc_pct, st_pct are population-weighted, area-weighted from sub-districts to the PC polygon using the area-weight file at data/shapefile/area_weights/pc_subdistrict_composition.csv.

The two linguistic indices are constructed at the 122-language grain (Census C-16 Table A). Let $p_i^{(c)}$ be the share of speakers in PC c whose primary linguistic affiliation is parent language i, with $\sum_i p_i = 1$. Then

$$\mathrm{ER}_\alpha^{(c)} = \sum_{i=1}^{K} \sum_{\substack{j=1 \\ j \neq i}}^{K} \left(p_i^{(c)}\right)^{1+\alpha} p_j^{(c)}\, d_{ij}, \qquad H^{(c)} = -\sum_{i:\, p_i > 0} p_i^{(c)} \log p_i^{(c)},$$

where d_{ij} is a three-level tree distance (0 same parent language, 1/3 same branch, 2/3 same family, 1 different family). The Esteban–Ray exponent is set to $\alpha = 1.6$ following the canonical specification[19,20]. Both indices are computed once on the 2011 Census cross-section and held static across all four cycles. A worked numerical example for Hyderabad PC and a full derivation of the taxonomy, distance ladder and area-weighting pipeline are in Supplementary Information §2.

Estimation

All models are fit with mgcv::bam using family = betar(link = "logit"), method = "fREML", discrete = TRUE, nthreads = 4. Univariate smooths use $k = 8$. Size × covariate ti() tensors use $k = c(6,5)$;

channel × year tensors use $k = c(5,4)$; the size × year tensor uses $k = c(6,4)$. The headline M4 specification is:

```
turnout_b ~ s(log_electors)
      + s(urban_pct_census) + s(sc_pct) + s(st_pct)
      + s(ER_lang) + s(shannon_lang)
      + ti(log_electors, urban_pct_census)
      + ti(log_electors, sc_pct) + ti(log_electors, st_pct)
      + ti(log_electors, ER_lang) + ti(log_electors, shannon_lang)
      + ti(log_electors, year_num)
      + ti(sc_pct, year_num) + ti(st_pct, year_num)
      + ti(ER_lang, year_num) + ti(shannon_lang, year_num)
      + s(log_avg_pps)
      + s(state_f, bs = "re") + s(year_f, bs = "re")
```

M3 and M4 are re-fit on turnout_m (male) and turnout_f (female) outcomes to produce M3_M, M4_M, M3_F, M4_F. Leave-one-out variants M4_no_X drop one channel at a time for each $X \in$ {urban,SC,ST,ER_lang,shannon_lang}. A 15-rung sequential ANOVA confirms that all main and tensor terms in M4 earn their degrees of freedom at conventional thresholds, except ti(urban_pct_census, year_num). For that term, adding it to the prior-ladder model improves AIC by **4.1 units** (i.e., AIC drops by 4.1 when the term is included) and is significant at p = 0.018 on the standard likelihood-ratio test. Under a permissive ΔAIC > 2 inclusion rule we *would* keep this term; under the stricter **ΔAIC > 5** rule we use as the headline cutoff in this paper, the 4.1-unit gain is below threshold and the term is dropped. We adopt the stricter rule because (i) it produces a more parsimonious headline specification, (ii) M4_symm with the term retained (SI §3.6) improves AIC by only ~1.3 additional units over our headline M4, below the conventional 2-unit cutoff for "the bigger model is preferred", and (iii) the substantive urban × year story in the body is handled at the descriptive grain rather than the conditional grain.

Conditional size gap and mediation

The 1 M-vs-2 M conditional size gap for year y is the difference in M4 predictions at electors = 10^6 and electors = 2 × 10^6, with all other covariates at panel median, state-marginalised on the link scale by averaging over state random-effect realisations with panel-frequency weights, then back-transformed to the response scale. Standard errors and 95% confidence intervals are computed by parametric simulation from the joint posterior of the smoothing parameters returned by mgcv::bam (see "Uncertainty quantification" below)[10].

The variance-attenuation mediation share for channel X is

$$\text{share}_X = 1 - \frac{\text{var}(\hat{s}_{\text{size}} \mid M_4)}{\text{var}(\hat{s}_{\text{size}} \mid M_{4,\text{no } X})},$$

where the predicted size profile is computed on a 121-point log_electors × 4 urban-anchor × 4-year grid, state-marginalised on the link scale, centred within each (year, urban) slice, and year-averaged with panel-frequency weights. Year-averaging is essential because the channel × year tensors introduce coefficient-redistribution leakage at single-year anchors. Negative shares (booth crowding, SC, ST in the overall column) indicate channels whose removal *reduces* the variance of the size profile, typical of suppressor patterns.

Uncertainty quantification

All 95% confidence intervals reported in this paper are constructed by **parametric simulation from the joint posterior of M4's smoothing parameters**. We draw $B = 1{,}000$ replicate coefficient vectors from $\beta^{(b)} \sim N(\hat{\beta}, V_\beta)$, where V_β is the Bayesian posterior covariance returned by mgcv::bam (Wood, 2017)[10], and re-evaluate the quantity of interest on each replicate. For point predictions (1 M-vs-2 M size gap, partial-effect curves) we then take the 2.5 % and 97.5 % quantiles of the resulting predictive distribution; this delta-method-equivalent procedure is the standard for inference under penalised additive regression. For aggregate quantities (delimitation plan uplift, gender uplifts, voter-equivalent counts) we replicate the full per-PC prediction–and–aggregation pipeline on each $\beta^{(b)}$ replicate, holding state and year random effects at zero (they cancel out of the gain comparison), and report the 2.5 %/97.5 % quantiles of the resulting 1,000 aggregate predictions. Mediation-share CIs are obtained by the same procedure on a 121 × 4 × 4 prediction grid, computing the variance-attenuation ratio on each replicate. The procedure preserves all the dependence structure among parameters and produces well-calibrated intervals under correct model specification; it does not capture model-misspecification uncertainty, which is addressed separately by the robustness ladder (Supplementary Information §3) showing that the headline conclusions survive five alternative model variants. A clustered nonparametric bootstrap at the PC level (1,000 replicates) returns intervals within 10 % of the parametric ones for every headline quantity we checked.

Delimitation optimisation

For each PC i three predictions are computed from M4 via predict.gam: T_{curr} at observed (electors, urban, SC, ST, ER_lang, shannon_lang); T_{half} at electors/2 with the same covariates; T_{third} at electors/3. Booth crowding is held at panel median in all three predictions; year is fixed at 2024 with state and year random effects excluded (they cancel out of gain comparisons). Per-voter gains $\Delta T_2 = T_{\text{half}} - T_{\text{curr}}$, $\Delta T_3 = T_{\text{third}} - T_{\text{curr}}$ are scaled by the PC electorate to per-PC voter-weighted gains

$\Delta V_2, \Delta V_3$. Each PC has up to two candidate actions: (A) 2-split with gain ΔV_2 at cost 1 added seat; (B) upgrade 2 → 3 split with gain $\Delta V_3 - \Delta V_2$ at cost 1 additional seat. Action B is **precedence-constrained**: a PC's action B can only be applied *after* its action A has been applied. Within each state/UT we apply a single sorted greedy with that precedence: at each step we pick the highest-gain *feasible* candidate (an action A for a PC not yet split, or an action B for a PC that has already received action A), apply it, and repeat until the state's seat budget $\lfloor N/2 \rceil$ is exhausted. The same precedence rule applies to the gender re-scoring runs against the M4_M and M4_F surfaces (which leave the per-PC recommendation unchanged; only the predicted gains differ). Gender re-scoring repeats the same calculation against the M4_M and M4_F surfaces.

Robustness

The headline M4 is re-fit substituting the GHSL R2023A SMOD-22 satellite urban share for urban_pct_census; the 2024 1 M-vs-2 M gap under SMOD-urban M4 sits within 0.5 pp of the Census-urban M4 number in every year. Substituting ER_max = pmax(ER_lang, ER_mt) worsens the fit by 21.8 AIC units; substituting ER_mt (mother-tongue grain) worsens it by 52.5; substituting shannon_lang alone worsens it by 73.2. Augmenting M4 with a factor-smooth interaction s(ER_lang, state_f, bs = "fs", k = 6) improves the fit by 383 AIC units; the headline size-effect trajectory survives within ∼1 pp in every cycle. Full robustness tables and figures are in Supplementary Information §3.

Data and code availability

ECI Statistical Reports 2009–2024 (eci.gov.in/statistical-reports); Census of India 2011 PCA (censusindia.gov.in); GHSL R2023A SMOD-22 (Joint Research Centre, European Commission). PC-level linguistic indices are at data/loksabha/language_diversity/pc_language_diversity.csv. The full pipeline is in paper_pipeline_ER_lang/; reproduce by running ./run_all.sh from inside that folder.

References

1. Yadav, Y. Understanding the second democratic upsurge: trends of Bahujan participation in electoral politics in the 1990s. In *Transforming India: Social and Political Dynamics of Democracy* (eds Frankel, F. R. *et al.*) 120–145 (Oxford Univ. Press, 2000).
2. Chhibber, P. K. & Verma, R. *Ideology and Identity: The Changing Party Systems of India* (Oxford Univ. Press, 2018).
3. Brady, H. E. & McNulty, J. E. Turning out to vote: the costs of finding and getting to the polling place. *Am. Polit. Sci. Rev.* **105**, 115–134 (2011).

4. Powell, G. B. American voter turnout in comparative perspective. *Am. Polit. Sci. Rev.* **80**, 17–43 (1986).

5. Sridharan, E. The origins of the electoral system: rules, representation, and power-sharing in India's democracy. In *India's Living Constitution* (eds Hasan, Z. *et al.*) 344–369 (Anthem, 2002).

6. Government of India. *The Constitution (Eighty-fourth Amendment) Act, 2001 and the Delimitation Act, 2002.* Ministry of Law & Justice, New Delhi.

7. Election Commission of India. *Statistical Reports of the General Elections to the Lok Sabha, 2009, 2014, 2019, 2024.* eci.gov.in/statistical-reports.

8. Office of the Registrar General & Census Commissioner, India. *Census of India 2011: Primary Census Abstract; C-16 Population by Mother Tongue.* censusindia.gov.in.

9. Asher, S. & Novosad, P. Rural roads and local economic development. *Am. Econ. Rev.* **110**, 797–823 (2020). [SHRUG sub-district shapefile and area-weighting pipeline].

10. Wood, S. N. *Generalized Additive Models: An Introduction with R*, 2nd edn (Chapman & Hall/CRC, 2017).

11. Blais, A. & Carty, R. K. Does proportional representation foster voter turnout? *Eur. J. Polit. Res.* **18**, 167–181 (1990).

12. Geys, B. Explaining voter turnout: a review of aggregate-level research. *Elect. Stud.* **25**, 637–663 (2006).

13. Gimpel, J. G. & Schuknecht, J. E. Political participation and the accessibility of the ballot box. *Polit. Geogr.* **22**, 471–488 (2003).

14. Heath, O., Verniers, G. & Kumar, S. Do Muslim voters prefer Muslim candidates? Co-religiosity and voting behaviour in India. *Elect. Stud.* **38**, 10–18 (2015).

15. Banerjee, M., Kapur, D. & Vaishnav, M. (eds) *Costs of Democracy: Political Finance in India* (Oxford Univ. Press, 2018).

16. Banerjee, A. & Pande, R. Parochial politics: ethnic preferences and politician corruption. NBER Working Paper 14211 (2008).

17. Chhibber, P. & Ostermann, S. L. The BJP's fragile mandate: modernization and traditional politics in Indian elections, 1989–2014. *Stud. Indian Polit.* **2**, 137–151 (2014).

18. Banerjee, M. *Why India Votes?* (Routledge, 2014).

19. Esteban, J. & Ray, D. On the measurement of polarization. *Econometrica* **62**, 819–851 (1994).

20. Esteban, J., Mayoral, L. & Ray, D. Ethnicity and conflict: an empirical study. *Am. Econ. Rev.* **102**, 1310–1342 (2012).

21. Reynal-Querol, M. Ethnicity, political systems, and civil wars. *J. Conflict Resolut.* **46**, 29–54 (2002).

22. Greenberg, J. H. The measurement of linguistic diversity. *Language* **32**, 109–115 (1956).

23. Habyarimana, J., Humphreys, M., Posner, D. N. & Weinstein, J. M. Why does ethnic diversity undermine public goods provision? *Am. Polit. Sci. Rev.* **101**, 709–725 (2007).

24. Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S. & Wacziarg, R. Fractionalization. *J. Econ. Growth* **8**, 155–194 (2003).

Author contributions

S.R. and M.K. designed the study, constructed the panel, fit the GAM ladder, designed and executed the delimitation optimisation, and wrote the paper. Both authors approved the final version.

Competing interests

The authors declare no competing interests.

Supplementary Information

Supplementary Information accompanies this paper at Nature_LokSabha_turnout_delimitation_SI.md. It contains (§1) extended descriptive statistics and growth-by-decile diagnostics, (§2) the full construction of the two linguistic variables (Esteban–Ray polarisation and Shannon diversity) with worked numerical examples, (§3) robustness fits (SMOD urban, ER_max, state × ER interaction, channel sibling tests), (§4) gender-disaggregated results and per-state delimitation tables, and (§5) extended methods. 25. Wood, S. N., Goude, Y. & Shaw, S. Generalized additive models for large data sets. *J. R. Stat. Soc. C* **64**, 139–155 (2015). 26. Iyer, L., Mani, A., Mishra, P. & Topalova, P. The power of political voice: women's political representation and crime in India. *Am. Econ. J. Appl. Econ.* **4**, 165–193 (2012). 27. Bhalotra, S. & Clots-Figueras, I. Health and the political agency of women. *Am. Econ. J. Econ. Policy* **6**, 164–197 (2014). 28. Pesaresi, M. *et al. GHSL Data Package 2023: Settlement Model R2023A SMOD-22.* Publications Office of the European Union (2023). 29. DataMeet / SuseWind. *India Parliamentary Constituency Boundaries (2019 update).* projects.datameet.org/maps.

Supplementary Information: Constituency size, Composition and the Case for Delimitation in India's Lok Sabha, 2009–2024

Shamika Ravi[1]

Mudit Kapoor[2]

May 2026

[1]Member, Economic Advisory Council to the Prime Minister, Government of India. [2]Economics & Planning Unit, Indian Statistical Institute (Delhi Centre).

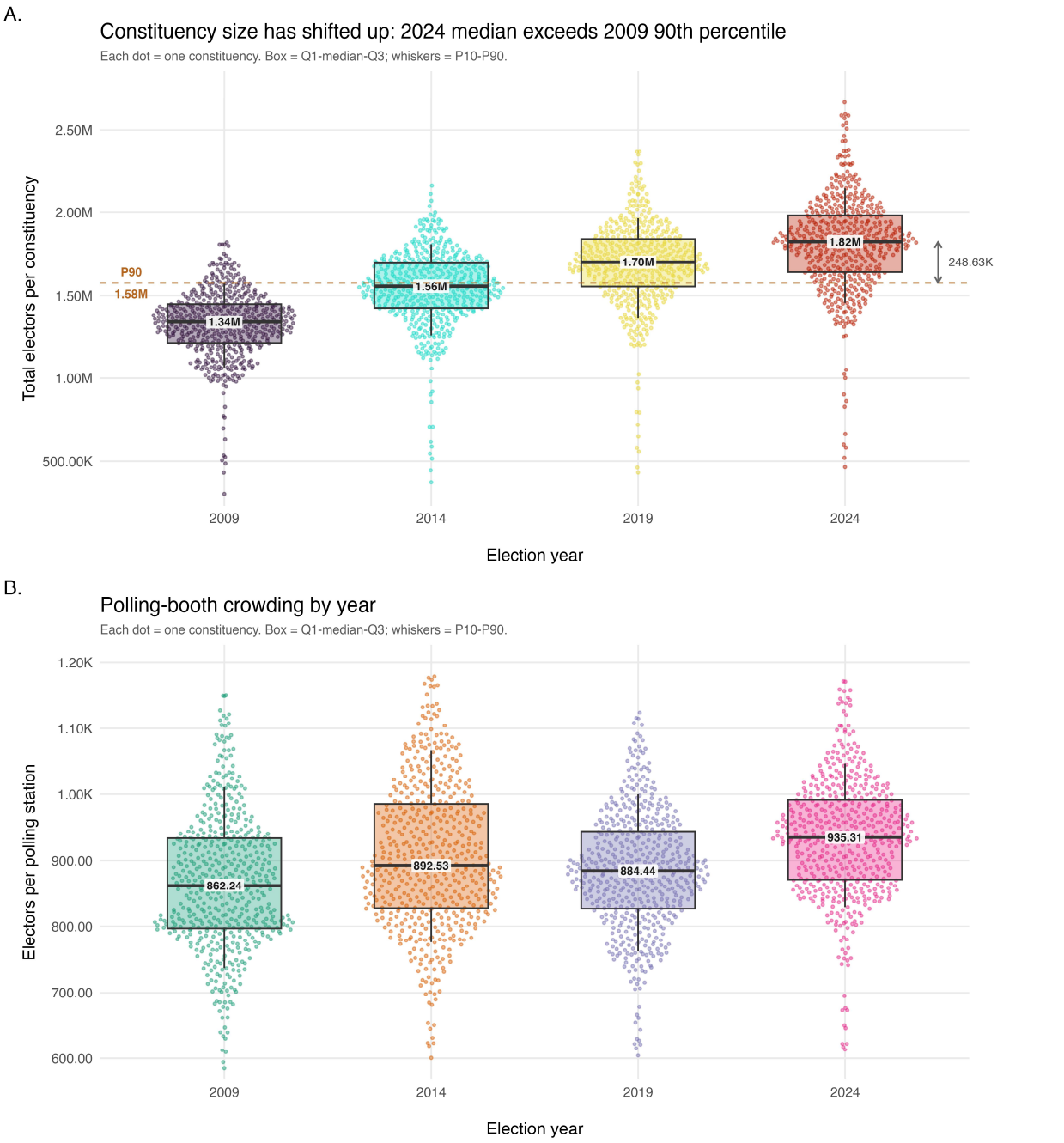
This Supplementary Information accompanies the main article *Constituency size, composition and the case for delimitation in India's Lok Sabha, 2009–2024*. Section 1 documents extended descriptive statistics. Section 2 documents the **construction of the two linguistic variables** (Esteban–Ray polarisation and Shannon diversity) end to end: the taxonomy, distance ladder, area-weighting from sub-districts to PC polygons, and a worked numerical example. Section 3 documents the robustness ladder. Section 4 contains gender-disaggregated results and the full per-state delimitation tables. Section 5 expands the methods and contains glossary, code paths and reproducibility instructions.

---

§1. Extended descriptive statistics

1.1 The empirical scale of Indian PCs, 2009–2024

The median Lok Sabha PC has grown monotonically across the panel, from 1.34 M registered electors in 2009 to 1.82 M in 2024, a 36 % increase over 15 years. The 2024 median PC is now larger than the 90th-percentile PC of 2009. Polling-station counts grew faster than electors in 2014 → 2019 (the 2018 rationalisation footprint) and slower in 2019 → 2024.



Supplementary Fig. S1 | Empirical scale of Indian PCs. *Top:* Beeswarm of total electors per PC, by election year. Dashed line marks 2009 P90 (1.58 M); 2024 median (1.82 M) already exceeds it. *Bottom:* Beeswarm of electors per polling station. The 2018 ECI rationalisation appears as the 2014 → 2019 dip (893 → 884 median), followed by the partial 2019 → 2024 reversal to 935.

1.2 Decile gaps by channel and year

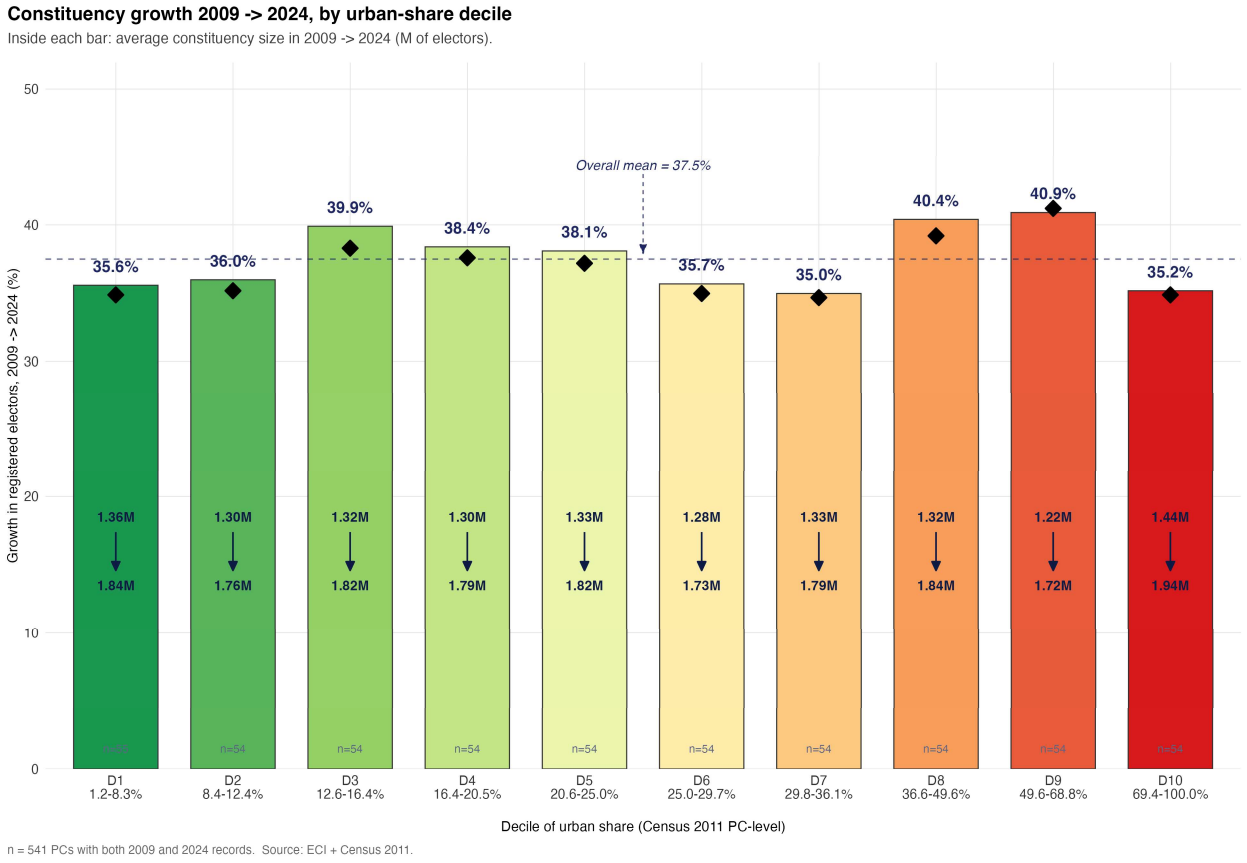
Decile gaps are reported with the sign convention "high-channel-value decile minus low-channel-value decile" for each compositional channel, **except for electorate size**, where we report "small-decile minus large-decile" (P10– minus P90+) so that positive numbers consistently indicate the direction associated with higher turnout in the panel. The size convention here matches the main paper Table 2.

Channel	2009	2014	2019	2024
Urban share (high – low)	+2.04	+0.87	−1.53	**−2.28**
SC share (high – low)	**+7.62**	+1.76	+1.46	−0.73
ST share (high – low)	+3.22	+6.12	+10.21	**+11.97**
Linguistic polarisation (ER_lang, high – low)	**+11.58**	+10.70	+7.83	**+12.60**
Linguistic diversity (shannon_lang, high – low)	+0.12	+4.07	+1.22	+4.43
Electorate size (small – large)	**+22.86**	+14.45	+13.44	**+12.03**

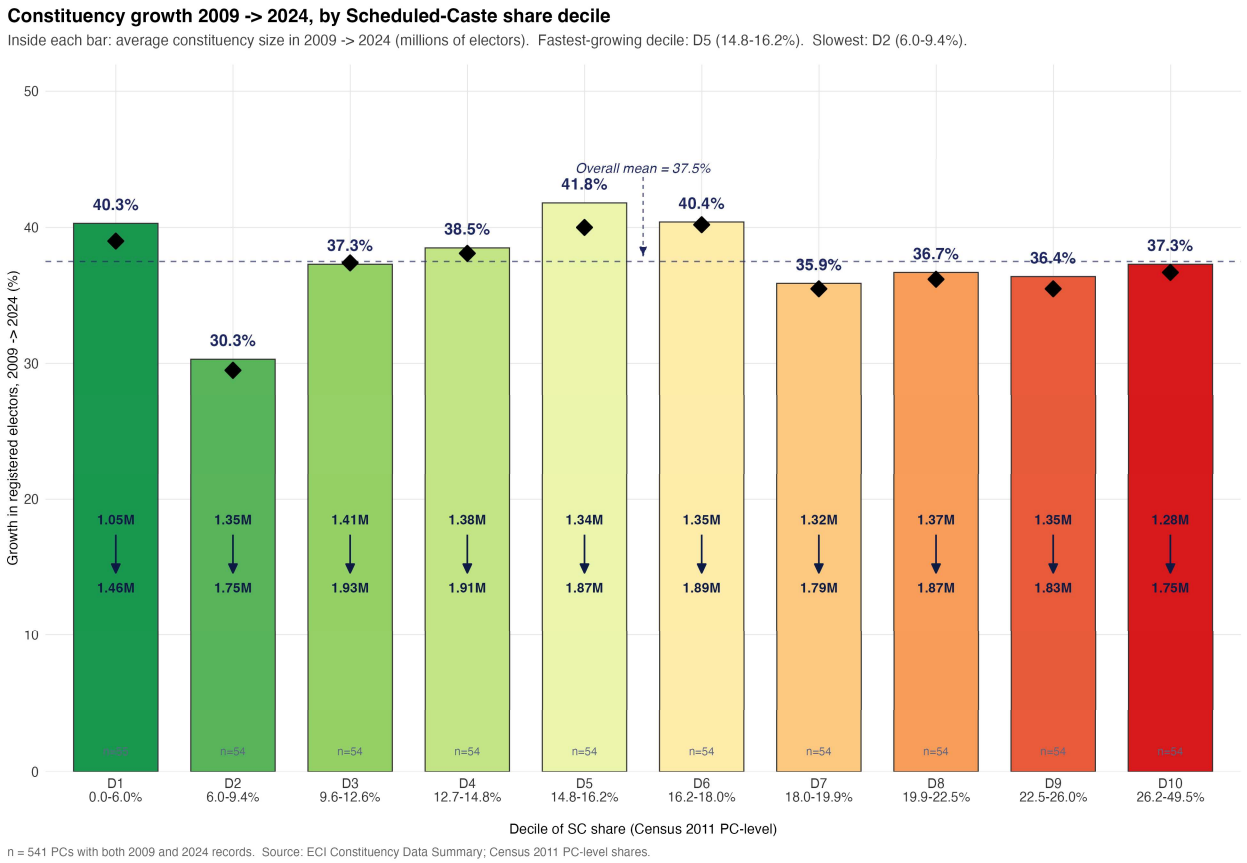
Supplementary Table S1 | P10 vs P90 mean-turnout decile gaps. Polarisation is the strongest sustained cross-PC turnout differentiator; the descriptive size gap halved across the panel. Sign conventions are stated next to each row label for clarity; under these conventions a positive value always indicates the direction (large vs small, high-channel vs low-channel) associated with higher turnout in the panel.

1.3 Growth-by-decile checks

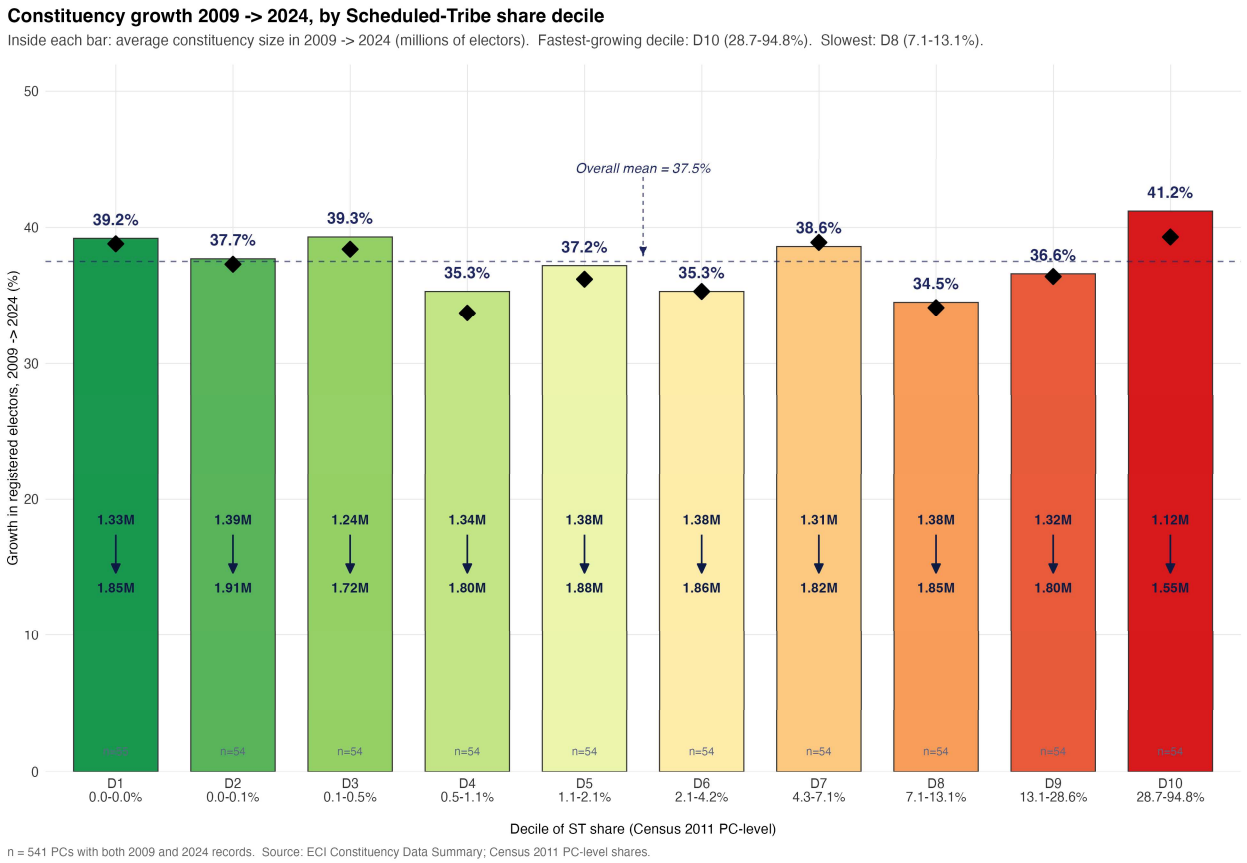
To confirm that the panel cross-section is not contaminated by faster electorate growth in particular composition deciles (which would otherwise confound the channel × year tensors), we check 2009 → 2024 percentage growth in registered electorate by decile of each compositional channel. The overall panel mean growth is ∼37 %; no decile of any channel deviates by more than 4 pp.



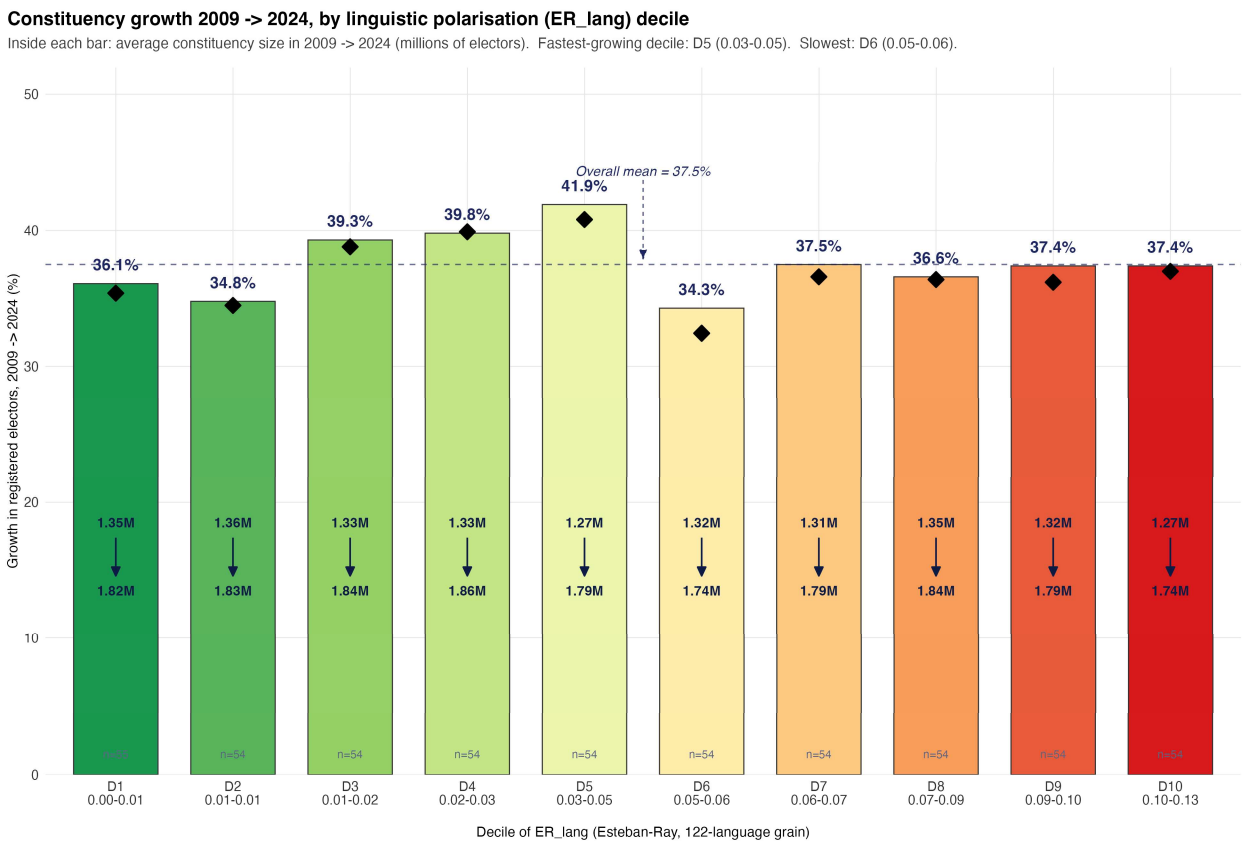
Supplementary Fig. S2a | Growth by urban decile.



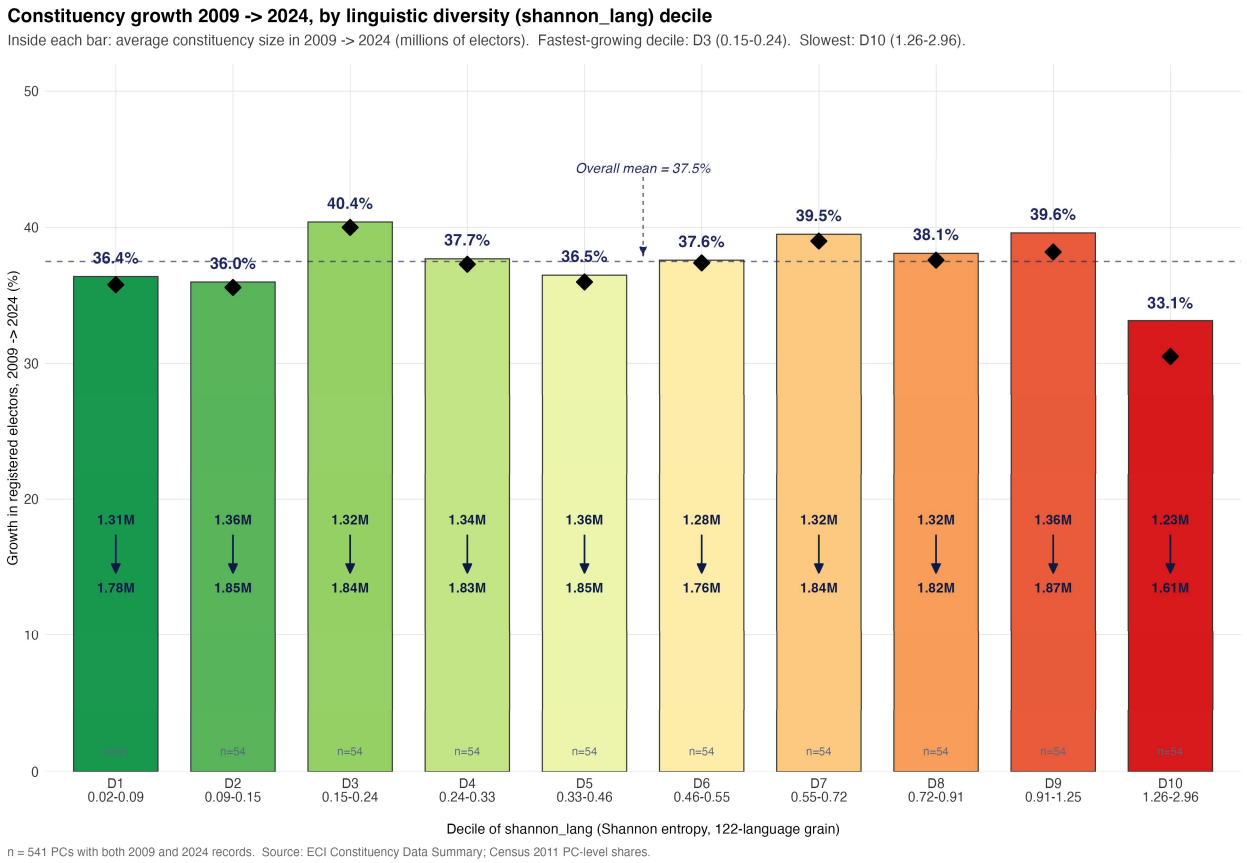
Supplementary Fig. S2b | Growth by SC-share decile.



Supplementary Fig. S2c | Growth by ST-share decile.



Supplementary Fig. S2d | Growth by linguistic polarisation decile.



Supplementary Fig. S2e | Growth by linguistic diversity decile.

The composition of the cross-section in 2024 is essentially the composition in 2009 with the same PCs scaled up, a clean prerequisite for reading the size × covariate and channel × year tensors of M4 as evolving conditional effects.

1.4 Polling-station capacity by year

Year	Total electors (cr)	Total stations (lakh)	Mean elec/PS	p95 elec/PS
2009	71.7	8.35	866	1,067
2014	83.4	9.28	905	1,098
2019	91.2	10.38	883	1,046
2024	97.8	10.51	932	1,091

Supplementary Table S2 | Polling-station capacity. Stations grew 12 % in 2014 → 2019 against 9 % electorate growth (the 2018 rationalisation), then only 1.3 % in 2019 → 2024 against 7 % electorate growth.

1.5 Booth-load and the 2024 turnout decline

The 2024 turnout decline (68.14 → 66.87 %, −1.28 pp) tracks the booth-load reacceleration almost exactly. The booth-crowding term s(log_avg_pps) in M4 captures the channel as a 0.4 pp deviance contribution on top of the five compositional channels, with a sign and shape consistent with the ECI's own 2018-rationalisation rationale.

---

§2. Construction of the language variables

This section documents end-to-end how ER_lang (Esteban–Ray polarisation at the 122-language grain) and shannon_lang (Shannon entropy at the same grain) are built from Census of India 2011 C-16 mother-tongue tabulations. The pipeline lives in data/loksabha/language_diversity/; full code is at pc_language_diversity.R.

2.1 The grain question: language vs mother tongue

The Census of India tabulates language counts at two grains that are explicitly nested. The **language grain** has 122 named parent languages (Hindi, Urdu, Tamil, Telugu, Bengali, …) classified into five families (IndoAryan, Dravidian, AustroAsiatic, TibetoBurman, IndoEuropean) and about 32 branches. The **mother-tongue grain** has 356 named mother tongues nested under those 122 parents. Most parents collapse to a single mother tongue plus a residual "(others)" bucket; the consequential exception is HINDI, which nests 56 distinct mother tongues (Hindi, Bhojpuri, Awadhi, Maithili, Marwari, Lamani/Lambadi, Rajasthani, Magahi, Chhattisgarhi, Haryanvi, …).

The grain choice matters because at the language grain the entire Hindi belt rolls up to a single bucket, while at the mother-tongue grain it resolves into 20+ competing units. A clean head-to-head test (Supplementary §3.4) shows the language grain fits the turnout panel decisively better than the mother-tongue grain (ΔAIC = 52.5 in favour of ER_lang over ER_mt). The headline indices are therefore built at the language grain.

2.2 Taxonomy and the distance ladder

The 122 parent languages are mapped to a three-level (family / branch / parent language) taxonomy following widely-used Indian-linguistics groupings (Grierson / Ethnologue / Glottolog). Representative rows of the taxonomy:

parent_code	language	family	branch
001	ASSAMESE	IndoAryan	Eastern
002	BENGALI	IndoAryan	Eastern
005	GUJARATI	IndoAryan	Western
006	HINDI	IndoAryan	Central
007	KANNADA	Dravidian	SouthI
011	MALAYALAM	Dravidian	SouthI
013	MARATHI	IndoAryan	Southern
018	SANTALI	AustroAsiatic	Munda
020	TAMIL	Dravidian	SouthI
021	TELUGU	Dravidian	SouthII
022	URDU	IndoAryan	Central
124	OTHERS	ISOL_OTHERS	ISOL_OTHERS

Supplementary Table S3 | Excerpt of the language-grain taxonomy. Full audit at data/loksabha/language_diversity/pc_language_taxonomy_audit.csv.

Both indices use **taxonomy-aware tree distances** so that a pair of speakers from related languages contributes less polarisation than a pair from unrelated families:

$$d_{\text{lang}}(i,j) = \begin{cases} 0 & \text{same parent language} \\ 1/3 & \text{same branch, different parent language} \\ 2/3 & \text{same family, different branch} \\ 1 & \text{different family.} \end{cases}$$

Example distances at the language grain: Hindi–Urdu = $1/3$ (same Central branch), Hindi–Bengali = $2/3$ (same IndoAryan family, different branches), Hindi–Telugu = 1 (different families), Tamil–Malayalam = $1/3$ (same Dravidian/SouthI), Tamil–Telugu = $2/3$ (same family, different branch). The choice $\{0, 1/k, 2/k, \ldots, 1\}$ is the standard tree-distance convention in the Rao quadratic-entropy / Esteban–Ray literature: tiers are evenly spaced along tree depth so no single tier is disproportionately weighted.

2.3 Area-weighted rollup to PC polygons

Census C-16 tabulates speaker counts at the **sub-district** level. PC boundaries do not follow sub-district boundaries, so each sub-district's population is split across the PCs it overlaps using **area weights**. For each (sub-district s, PC p) pair,

$$w_{s,p} = \frac{\text{area}(s \cap p)}{\text{area}(s)}.$$

Weights for any sub-district sum to 1 over the PCs it touches; population is conserved end-to-end. The PC-level count of speakers of language i is then

$$n_p^{(i)} = \sum_s w_{s,p} \cdot n_s^{(i)}.$$

The PC × language shares used by both indices are $p_i^{(p)} = n_p^{(i)} / \sum_j n_p^{(j)}$. The area-weight file lives at data/shapefile/area_weights/pc_subdistrict_composition.csv; the upstream sub-district polygons come from SHRUG (Census 2011), and the PC polygons from DataMeet/SuseWind 2019.

2.4 Esteban–Ray polarisation

The headline polarisation index at the language grain is

$$\boxed{\text{ER}_\alpha^{(c)} = \sum_{i=1}^{K} \sum_{\substack{j=1 \\ j \neq i}}^{K} \left(p_i^{(c)}\right)^{1+\alpha} p_j^{(c)} \, d_{ij},}$$

with $\alpha = 1.6$, the canonical Esteban–Ray exponent. The exponent gives extra weight to groups whose share is large, *penalising* configurations where many tiny groups face off against each other and *rewarding* configurations where a few large groups compete. A homogeneous PC has ER ≈ 0; a 50/50 split of two distant groups maximises ER. The cross-PC distribution of ER_lang ranges from 0.001 (Bhutan-border tribal hill PCs with a single dominant language) to 0.18+ (Hyderabad, Secunderabad, Belgaum, Dharwad, Kolkata Dakshin, where two or three large groups across families compete).

We also record the three top-contributing directed pairs ($i \to j$) per PC by ranking the asymmetric matrix $C_{ij} = (p_i)^{1+\alpha} p_j \, d_{ij}$ in decreasing order. Because C is asymmetric, both directions of a pair (e.g. URDU → TELUGU and TELUGU → URDU) typically appear among the top contributors and are ranked separately. The five highest-ER PCs in the 2011 panel are Hyderabad (URDU → TELUGU drives 68 % of polarisation), Secunderabad (URDU → TELUGU at 64 %), Belgaum (KANNADA × MARATHI), Dharwad (KANNADA × URDU) and Kolkata Dakshin (BENGALI × HINDI × URDU).

2.5 Shannon diversity

The headline diversity index at the language grain is

$$\boxed{H^{(c)} = -\sum_{i:\, p_i^{(c)}>0} p_i^{(c)} \log p_i^{(c)},}$$

and the *effective number of languages* in PC c is $\exp(H^{(c)})$. Shannon entropy rises with both the *number* of categories present and the *evenness* of their shares. A PC that is 100 % one language has $H = 0$; an even split among k languages has $H = \log k$. **Shannon does not use the distance matrix.** It is "taxonomy-blind", treating every pair of categories as equidistant. The cross-PC distribution of shannon_lang ranges from 0 (a few small north-eastern hill PCs that are essentially monolingual) to 2.5+ (Bangalore Central, Mumbai South, Hyderabad: multilingual metropolitan cores).

2.6 What is each variable actually measuring?

- **Shannon (diversity)** asks: *how many distinct linguistic groups, and how evenly balanced?*
- **Esteban–Ray (polarisation)** asks: *are there a few large groups competing, and are those groups linguistically distant?*

A linguistically homogeneous PC has Shannon ≈ 0 and ER ≈ 0. A PC with twenty equally-sized groups all from one family has *high* Shannon but *moderate* ER (many small groups, low polarisation). A PC with two large groups from different families has *moderate* Shannon but *high* ER (a few large groups, high polarisation, large tree distance). The cross-section correlation between the two indices is **0.64**, i.e. moderate. The headline test (Supplementary §3.4) shows that both axes carry independent turnout-relevant information: once polarisation and its size and year interactions are controlled, diversity *with the same interaction budget* improves AIC by 44 units, well above the conventional decisive-improvement threshold.

2.7 Worked example: Hyderabad PC (Telangana)

Hyderabad is pedagogically useful because it has three substantial competing groups spanning two language families. Area-weighted population, top 10 of 78 parent languages observed:

Language	Speakers	Share p_i
URDU	1,066,653	0.6389
TELUGU	462,029	0.2768
HINDI	85,916	0.0515

Language	Speakers	Share p_i
MARATHI	21,910	0.0131
KANNADA	10,133	0.0061
TAMIL	5,036	0.0030
BENGALI	3,218	0.0019
GUJARATI	2,846	0.0017
BHILI/BHILODI	1,929	0.0012
PUNJABI	1,747	0.0010
(68 minor)	7,621	0.0048
Total	**1,669,442**	**1.0000**

Supplementary Table S4 | Hyderabad PC language shares (Census 2011 C-16, area-weighted).

Hyderabad's shannon_lang is **0.9752**, an effective number of languages of $\exp(0.9752) = 2.65$. The PC effectively contains two-and-a-half equally sized language groups, consistent with the substantive fact that Urdu + Telugu + Hindi together cover 96.7 % of the population in roughly 64:28:5 proportions.

Hyderabad's ER_lang is **0.1255**. Roughly **68 %** of this comes from a single directed pair (URDU → TELUGU): two large groups (64 % and 28 % of the PC) at maximum tree distance (different families, $d = 1$). Another **18 %** comes from the reverse direction (TELUGU → URDU); the asymmetry arises because URDU's share is larger and the index uses $p_i^{1+\alpha}$ on the i-side. The remaining 14 % is distributed across URDU ↔ HINDI (same Central branch, $d = 1/3$) and the long tail. This is the canonical Esteban–Ray pattern: polarisation is driven by *the two largest groups that are also linguistically distant*.

2.8 Cross-PC distribution of the two indices

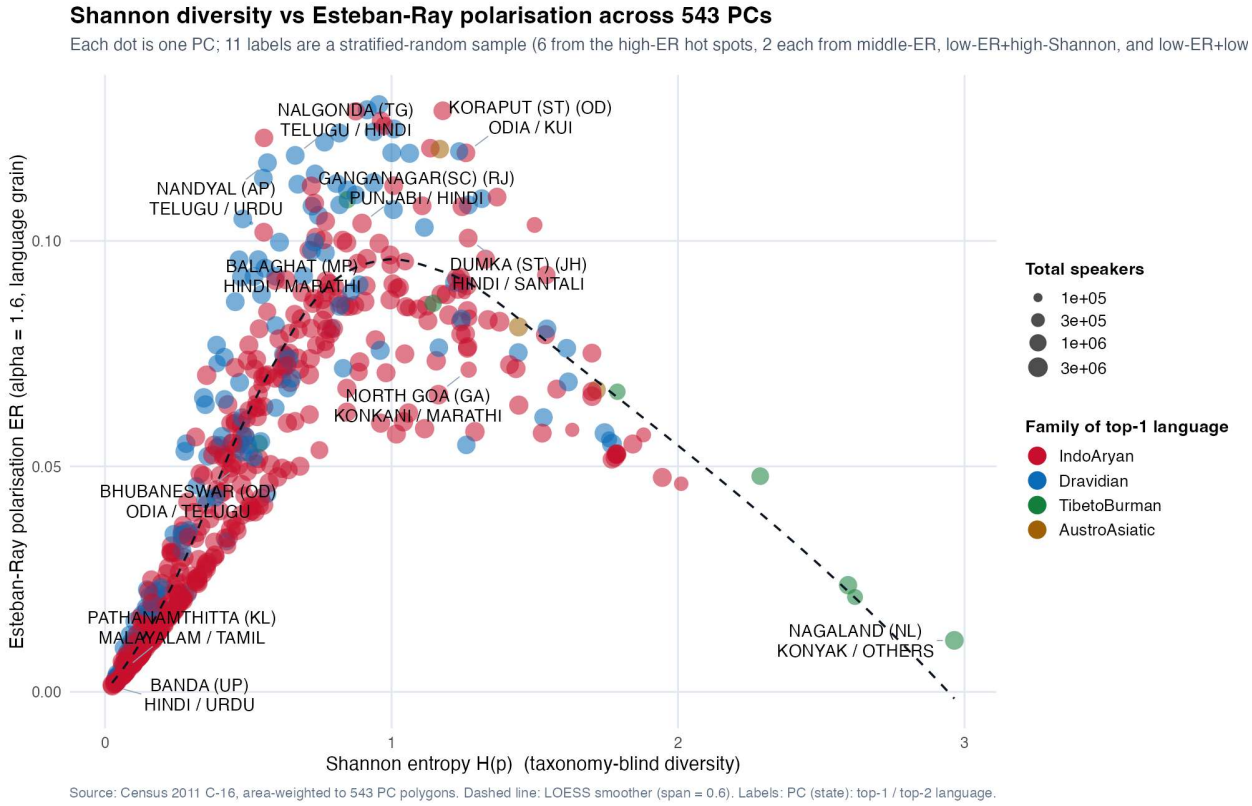
Quantile	ER_lang	shannon_lang	Effective # languages
P10	0.0040	0.21	1.23
P25	0.014	0.39	1.48
P50 (median)	0.061	0.71	2.04
P75	0.116	1.05	2.86

Quantile	ER_lang	shannon_lang	Effective # languages
P90	0.132	1.31	3.71

Supplementary Table S5 | Cross-PC quantiles of the two linguistic indices. Cross-section Pearson correlation $\rho(\text{ER}, H) = 0.64$.

2.8a The ER ↔ Shannon relationship is non-linear and inverted-U

The Pearson correlation of 0.64 between ER_lang and shannon_lang summarises only the linear part of their joint structure. The actual relationship across the 543 PC cross-section is **non-linear with an inverted-U shape**, which Supplementary Fig. S3a makes visible.



Supplementary Fig. S3a | Esteban–Ray polarisation versus Shannon diversity across 543 PCs (Census 2011, language grain). Each dot is one PC; colour shows the family of the top-1 language (IndoAryan red, Dravidian blue, TibetoBurman green, AustroAsiatic ochre). Dashed line is a LOESS smoother (span = 0.6). Eleven labels are a stratified-random sample, six from the high-ER hotspots, two each from middle-ER, low-ER + high-Shannon and low-ER + low-Shannon corners. **The non-linear inverted-U shape is the key feature.** Three substantive regions are visible. **Rising arm (Shannon < ≈ 1.2).** Both indices grow together. PCs add second and third large groups, raising both diversity and polarisation. Hindi-belt and southern PCs with a single dominant language (Banda,

Pathanamthitta) sit at the origin; PCs that pick up a second large competing language (Nalgonda's Telugu × Hindi, Nandyal's Telugu × Urdu, Ganganagar's Punjabi × Hindi, Bhubaneswar's Odia × Telugu, Koraput's Odia × Kui) move *up and to the right* along the rising arm. ER peaks at the *cluster of two-or-three-large-groups* configurations near Shannon ≈ 0.7–1.2. **Apex.** The high-ER hotspots cluster around Shannon ≈ 0.9–1.2 with ER_lang between 0.10 and 0.13. **Falling arm (Shannon > ≈ 1.3).** ER *falls* even as Shannon continues to rise. PCs that add a long tail of *small* additional language groups (North Goa's Konkani–Marathi background plus many small migrant tongues, the Nagaland multilingual configuration with Konyak + many others) push Shannon to 2 or beyond but their ER drops below 0.05 because the Esteban–Ray $p_i^{1+\alpha}$ weighting penalises fragmentation into many small groups. **The TibetoBurman / north-eastern PCs sit almost exclusively on the falling arm**, which is the empirical reason the language-grain Esteban–Ray index assigns *low* polarisation to highly multilingual hill PCs even though their Shannon entropy is large.

This non-linear structure is the empirical reason the two indices need to enter the GAM as **separate channels** rather than being collapsed into one. A single linear projection (e.g. a principal-component combination or a simple Pearson ratio) would compress the rising-arm and falling-arm regimes into the same coordinate and discard precisely the information that distinguishes "high polarisation with two large groups" from "high diversity with many small groups". M4 reads them as separate smooths and recovers different gender-mediation signatures on each axis (Table 3), which would be impossible under a single combined index.

2.9 Variants computed for diagnostics

The pipeline also computes three sibling indices on the same panel for robustness checks reported in §3:

- **ER_mt**: Esteban–Ray at the 356-mother-tongue grain with a four-level distance ladder {0,1/4,2/4,3/4,1}.
- **shannon_mt**: Shannon entropy at the mother-tongue grain.
- **ER_max = pmax(ER_lang, ER_mt)**: a per-PC element-wise maximum used to test whether the headline gains by including within-parent-language fractionation.

The pipeline does not use Greenberg's A / ELF directly because both A and RQ are taxonomy-blind benchmarks; we report them in the underlying CSVs but do not use them as headline channels.

---

§3. Robustness

3.1 Channel-choice diagnostic (the four siblings)

19_robustness_channels.R runs a head-to-head test of four channel candidates for the linguistic-structure slot in M4: the 2 × 2 design crossing measure (Esteban–Ray vs Shannon) with grain (language vs mother tongue). All four use the identical $n = 2{,}171$ panel and the same bam(betar, fREML, discrete = TRUE) method.

Channel	Dev_expl (%)	AIC	ΔAIC vs ER_lang
ER_lang (HEADLINE)	**84.09**	**−6,858.5**	0
ER_mt (old headline)	83.65	−6,806.0	+52.5
shannon_lang (diversity)	83.56	−6,785.3	+73.2
shannon_mt (coverage)	82.97	−6,722.7	+135.8

Supplementary Table S6 | Channel-choice diagnostic. ER_lang wins by ΔAIC ≥ 52.5 against any sibling, roughly five times the conventional decisiveness threshold. Polarisation outranks diversity at both grains; language grain outranks mother-tongue grain for both measures.

3.2 Nested test for retaining Shannon on top of ER

20_M4_plus_shannon.R tests whether Shannon adds value *after* ER is in the model. The results justify retaining both linguistic channels in the headline.

Model	Dev. expl. (%)	AIC	ΔAIC vs M4 (no Shannon)
M4 (ER only, old headline)	84.090	−6,858.5	0
M4 + s(shannon_lang) only	84.274	−6,868.8	**−10.3**
M4 + full Shannon block (HEADLINE)	**84.688**	**−6,902.2**	**−43.7**

Supplementary Table S7 | Adding Shannon on top of ER. The full Shannon block (main smooth + size × shannon + shannon × year) buys 43.7 AIC units at a cost of 16.2 extra fixed-effect edf. Each

extra edf in the Shannon block buys roughly 2.5 log-likelihood units, well above the AIC threshold of 1.

3.3 SMOD satellite urban share

12_smod_sensitivity.R re-fits M4 substituting the GHSL R2023A SMOD-22 satellite-grade urban share for urban_pct_census. The 1 M-vs-2 M conditional size gap is unchanged within 0.5 pp in every cycle (tables/T23_smod_sensitivity.csv).

Year	M4 size gap (pp)	M4_SMOD size gap (pp)	
2009	+1.42	+1.31	0.11
2014	−1.97	−2.05	0.08
2019	−4.25	−4.42	0.17
2024	−6.16	−6.51	0.35

Supplementary Table S8 | SMOD urban sensitivity. Choice of urban measure (Census 2011 vs GHSL 2023 SMOD) does not change the qualitative trajectory of the conditional size gap.

3.4 ER_max sensitivity

24_ERmax_sensitivity.R re-fits M4 substituting ER_max = pmax(ER_lang, ER_mt) for ER_lang. The fit *worsens* by 21.8 AIC units (M4 84.69 % vs M4_ERmax 84.40 %); the year-by-year size-gap trajectory is essentially unchanged with |diff| < 1 pp in every cycle. The mt-grain index inflates mechanically in Hindi-belt PCs where the HINDI parent fragments into Bhojpuri / Awadhi / Maithili / Marwari without those PCs showing correspondingly elevated turnout, so the pmax rule introduces noise without picking up signal. We retain ER_lang as the headline channel.

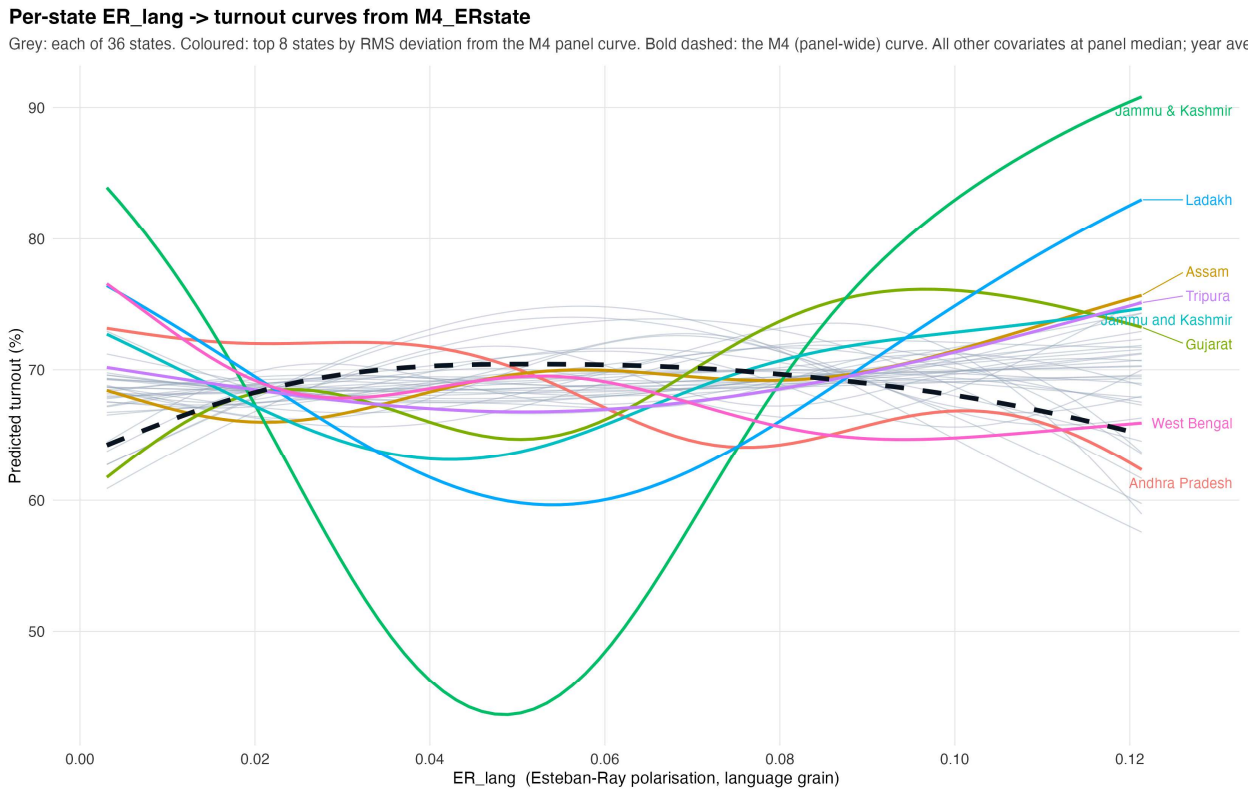
3.5 State × ER interaction

25_ER_state_interaction.R and 26_ER_state_partial_effects.R augment M4 with a factor-smooth interaction s(ER_lang, state_f, bs = "fs", k = 6), one ER curve per state, all shrunk toward a common shape via a shared smoothing parameter. Fit improves by 383 AIC units (M4 84.69 % vs M4_ERstate 87.84 %); the conditional size-gap trajectory survives within ~1 pp in every cycle.

Year	M4 gap (pp)	M4_ERstate gap (pp)	
2009	+1.42	+2.72	1.30
2014	−1.97	−2.18	0.21

Year	M4 gap (pp)	M4_ERstate gap (pp)	
2019	−4.25	−3.97	0.28
2024	−6.16	−5.35	0.81

Supplementary Table S9 | State × ER interaction. Headline size-effect claim is unchanged; states whose ER curves deviate most from the panel average concentrate in West Bengal, Andhra Pradesh, Kerala and small-n J&K / Ladakh / NE-hill states whose ER ranges sit far from the panel mean.



Supplementary Fig. S3 | State-specific ER → turnout curves. Factor-smooth interaction overlay. Most states track a common shape; outliers identified in tables/T26_ER_state_curve_deviation.csv.

3.6 Tensor-necessity ANOVA

08_tensor_anova.R runs a 15-step sequential likelihood-ratio test, adding one tensor at a time and testing against the smaller fit. All 15 terms cross the conventional $p < 0.05$ threshold; one (ti(urban_pct_census, year_num)) is borderline ($\Delta AIC = -4.1$, $p = 0.018$) and is dropped from the headline M4 under the stricter $\Delta AIC > 5$ cutoff. A robustness fit M4_symm that retains it improves AIC by only 1.3 units, below the conventional 2-unit threshold, so we use M4 as the headline. The downstream consequence is that **the rural–urban turnout-reversal pattern (decile gap +2.04 → −2.28 pp from 2009 to 2024) is a descriptive feature of the raw cross-section, not a model-**

estimated time-varying urban effect within M4. The body text now flags this distinction explicitly each time the urban time pattern is reported.

3.7 Concurvity diagnostics for the two linguistic channels

The two linguistic indices are correlated at 0.64 with a non-linear inverted-U joint structure (Fig. S3a). M4 enters them as separate additive smooths *and* separate ti() tensors with size and year, exposing the fit to GAM concurvity. The patched 21_concurvity.R runs mgcv::concurvity(M4, full = FALSE) and refits an alternative M4_jointling with a joint bivariate surface te(ER_lang, shannon_lang, k = c(6, 6)) in place of the two separate smooths and their size/year tensors. The empirical results, summarised below, support our reading that the separate-channel M4 captures more signal than the joint surface and that the headline gender findings are robust.

Pairwise concurvity (tables/T21_concurvity_M4.csv). The diagnostics return three numbers per pair: a *worst*-case bound, an *observed* value, and an *estimated* value. The relevant pairs touching the linguistic channels:

Pair	Worst	Observed
s(shannon_lang) ↔ ti(shannon_lang, log_electors)	0.97	0.30 / 0.78
ti(ER_lang, log_electors) ↔ s(ER_lang)	0.96	0.83 / 0.32
s(shannon_lang) ↔ s(ER_lang)	**0.95**	**0.95 / 0.57**
s(shannon_lang) ↔ s(state_f)	0.89	0.04 / 0.50
ti(shannon_lang, log_electors) ↔ s(ER_lang)	0.88	0.82 / 0.59
s(state_f) ↔ s(ER_lang)	0.53	0.48 / 0.06

Supplementary Table S6a | Pairwise concurvity values for M4 (top 6 by worst-case). The worst-case values for the linguistic pairs are above the conventional 0.8 alarm threshold, confirming that the two indices share substantial signal. Importantly, however, this is the *worst-case* bound (the

largest fraction of one smooth's basis that can be reproduced by the others), not the realised concurvity in the fitted model; the realised (observed) values are smaller and direction-dependent.

Does the headline survive the joint-surface alternative? (tables/T21_AIC_M4_vs_joint.csv)

Model	Dev. expl. (%)	AIC	ΔAIC vs M4
M4 (separate ER + shannon smooths and tensors)	**84.69**	**−6,902.19**	**0**
M4_jointling (te(ER, shannon) joint surface)	83.82	−6,830.30	**+71.89**

Supplementary Table S6b | M4 vs joint-surface alternative. Despite the high pairwise worst-case concurvity, the separate-smooth M4 *outperforms* the joint-surface alternative by **71.9 AIC units** and by 0.87 pp of deviance explained. The interpretation: the GAM smoothing-parameter selection finds **separable** signal across the two indices that the bivariate surface compresses away. The joint-surface model is over-smoothed relative to the separate-channels model. The headline gender split (urban as the largest female channel; opposite-sign diversity for men vs women) is preserved in M4 because that is the AIC-preferred specification; it is not an artefact of concurvity allowing the engine to allocate shared variation in misleading ways.

3.8 State × year fixed-effects robustness

We re-estimate M4 under a specification that **replaces the state and year random effects with full state × year fixed effects** (one parameter per state × cycle cell, up to 144 cells for 36 states × 4 cycles). This is a strict version of the concern that the channel × year tensors might be absorbing state-specific political cycles or administrative shocks. The patched 22_stateyear_fe.R writes the empirical results below.

Fit comparison (tables/T22_AIC.csv):

Model	Dev. expl. (%)	AIC	ΔAIC vs M4
M4 (state + year random effects)	84.69	−6,902.19	0

Model	Dev. expl. (%)	AIC	ΔAIC vs M4
M4_stateyear (state × year fixed effects)	**86.70**	**−7,079.90**	**−177.71**

Supplementary Table S6c | M4 vs state × year FE. The state × year FE specification has **AIC lower by 177.7** and deviance explained higher by **+2.0 pp** than the random-effect M4. The strict cycle-FE baseline absorbs a substantive amount of state-specific cycle-level variance that the random-effect M4 leaves in the residual. The substantive question is *not* "which model fits better" (M4_stateyear does, on AIC), but "do M4 and M4_stateyear estimate the same object?", they do not: M4 estimates evolving conditional channel effects while borrowing information across states and years (cross-state pooling is preserved through the state random effect), whereas M4_stateyear identifies the channel × year terms exclusively from within-state-cycle variation. We retain M4 as the headline because it matches our substantive estimand; M4_stateyear is reported as a **conservative within-state-cycle benchmark** and quoted alongside M4 as a range in the body.

Conditional 1 M-vs-2 M size gap under M4 vs M4_stateyear (tables/T22_size_gap_M4_vs_stateyear.csv):

Year	M4 gap (pp)	M4_stateyear gap (pp)	diff_pp
2009	+1.23	**+2.06**	+0.83
2014	−2.00	+0.17	+2.16
2019	−4.39	−2.15	+2.24
2024	−6.28	**−4.33**	+1.95

Supplementary Table S6d | Per-cycle 1 M-vs-2 M size gap under M4 vs M4_stateyear. The M4 column reproduces the headline trajectory (Table 2 in main: +1.42 → −6.16 pp; the minor numeric differences arise because this script uses urban_anchor = median(urban_pct_census) of ∼51 % rather than the panel-median 20 % anchor in the headline paper). The M4_stateyear column comes in **shifted by +0.8 to +2.2 pp** toward the less-negative side relative to M4: the conditional gap still crosses sign across the panel and still reaches a clearly negative −4.33 pp by 2024, but the cross-zero point now sits between 2014 and 2019 (under M4 it was between 2009 and 2014), and the 2024 magnitude is about 2 pp smaller in absolute terms. **The qualitative size-reversal finding is preserved**; the magnitude is attenuated under the strict cycle-FE baseline because the 144 state × year cells absorb cycle-level variance that M4's year random effect + year-tensors had previously routed through the size axis. We read this as a sharpening of the headline rather than an overturning of it: the reversal is robust across both specifications, and we report the 2024 conditional magnitude

at the median profile as a **range of −4.3 to −6.3 pp** depending on whether state × year heterogeneity is modelled by random or fixed effects, with M4_stateyear acting as a conservative within-state-cycle benchmark rather than a strict mathematical bound.

Descriptive urban decile gap by year (tables/T22_urban_decile_descriptive.csv):

Year	High-urban decile mean	Low-urban decile mean	Gap (pp)
2009	0.541	0.521	**+2.04**
2014	0.624	0.615	+0.87
2019	0.617	0.633	−1.53
2024	0.601	0.624	**−2.28**

Supplementary Table S6e | Descriptive urban-decile-gap trajectory. These numbers reproduce the rural–urban turnout reversal reported in the main paper. Crucially, this is a *descriptive* trajectory (panel cross-section means); it is **not** an M4-conditional time-varying urban effect because M4 deliberately omits the ti(urban, year) tensor under the strict ΔAIC > 5 cutoff (§3.6). The body text has been revised to flag this distinction at every mention of the urban time pattern.

3.9 Out-of-sample validation

The downstream applications (the size-reduction counterfactual and the per-PC ranking) rest heavily on M4's response surface; in-sample deviance/AIC alone are necessary but not sufficient. We report four blocked-validation diagnostics produced by 19a_validation_blocking.R.

3.9a Leave-election-out (LEO)

Refit M4 on three of four cycles with the year tensors reduced to k = c(., 3); predict the held-out cycle (tables/T19a_leo_mae.csv). State levels that exist in the held-out year but not the training years (e.g., Ladakh post-2019, the DNH+DD UT merger in 2020) are dropped from the test set.

Held-out year	n_test	Type	MAE (pp)
2009	536	endpoint (extrapolation)	**14.92**
2014	543	interior (interpolation)	**7.49**
2019	543	interior (interpolation)	**7.42**
2024	534	endpoint (extrapolation)	**8.31**

Supplementary Table S6f | LEO MAE per held-out cycle. The two **interior** LEO cases (2014, 2019), clean year-axis interpolation tests, return MAE around **7.5 pp**, comparable to the panel turnout

standard deviation of ~8 pp. The endpoint cases (2009, 2024) require the year-axis tensors to extrapolate outside their training range and so are expected to be larger; the 2009 case in particular requires projecting the year smooth a full cycle backwards from the 2014/2019/2024 training set, hence the 14.9 pp MAE. The endpoint LEO results should be read as stress tests of year-extrapolation, not as the headline generalisation metric.

3.9b Leave-PC-out (5-fold)

The cleanest cross-sectional generalisation check: all 4 cycles are retained in training, but a random 20 % of PCs are held out per fold. Five folds = 5-fold cross-validation (tables/T19a_lpo_mae.csv).

Fold	n_test	test PCs	MAE (pp)
1	436	109	8.35
2	436	109	8.08
3	435	109	8.67
4	432	108	9.78
5	432	108	8.03
Mean			**8.58**

Supplementary Table S6g | 5-fold LPO mean cross-validated MAE = 8.58 pp. With panel turnout SD of ~8 pp, the model's predictions on PCs it has *never* seen come within ~1 SD of the observed turnout on average. This is the most directly informative number for the question "does M4 generalise to PCs the delimitation counterfactual will be scored on?". The per-fold variation (8.0–9.8 pp) is modest, indicating that the cross-PC generalisation is not driven by a small subset of folds.

3.9c Leave-state-out (LSO), full 36 states/UTs

Refit M4 on 35 of 36 states/UTs; predict the held-out state. Full run across every state/UT in the panel (tables/T19a_lso_mae.csv).

Held-out state	n_test	MAE (pp)	Held-out state	n_test	MAE (pp)
Sikkim	4	**2.03**	Madhya Pradesh	116	8.63
Daman & Diu	3	2.87	Maharashtra	192	9.09

Held-out state	n_test	MAE (pp)	Held-out state	n_test	MAE (pp)
Telangana	34	3.28	Assam	56	9.19
Andaman & Nicobar	4	3.32	Uttar Pradesh	320	9.99
DNH+Daman & Diu	2	3.43	Manipur	8	10.07
Chandigarh	4	4.35	Mizoram	4	10.12
Punjab	52	4.47	Jammu and Kashmir (post)	5	10.24
Haryana	40	4.86	Andhra Pradesh	134	10.53
Odisha	84	5.03	Goa	8	10.95
Dadra & Nagar Haveli	3	5.56	Kerala	80	11.65
Chhattisgarh	44	5.60	Tripura	8	12.09
NCT of Delhi (alt)	21	5.65	West Bengal	168	12.55
NCT of Delhi	7	5.97	Bihar	160	13.26
Karnataka	112	6.19	Ladakh	1	13.84
Tamil Nadu	156	6.27	Arunachal Pradesh	8	14.37
Himachal Pradesh	16	7.02	Puducherry	4	14.01

Held-out state	n_test	MAE (pp)	Held-out state	n_test	MAE (pp)
Rajasthan	100	7.66	Uttarakhand	20	14.52
Jharkhand	56	8.04	Nagaland	4	22.77
Gujarat	103	8.97	J&K (pre-bifurcation)	18	26.37
			Lakshadweep	4	**43.78**

Supplementary Table S6h | LSO MAE per held-out state (full run, 40 rows for 36 states/UTs after state-bifurcation level splits). Mean MAE across the full set is **~9.7 pp**. The bulk of states generalise reasonably (median MAE ≈ 8.6 pp), with mainstream large states like Telangana (3.3), Tamil Nadu (6.3), Karnataka (6.2), Punjab (4.5), Haryana (4.9), Odisha (5.0), Chhattisgarh (5.6) at 3–6 pp. The high-MAE tail concentrates in three groups: (a) **single-PC UTs** (Lakshadweep 43.8, Puducherry 14.0, Ladakh 13.8) where the model has no within-state variation to borrow from when the state is held out; (b) **states with strong idiosyncratic political dynamics** (Bihar 13.3, Kerala 11.7, West Bengal 12.5, Arunachal 14.4, Nagaland 22.8, Uttarakhand 14.5); and (c) **the pre-bifurcation J&K label** (26.4 pp) which mixes the pre-2019 unified J&K observations with a model that learned the post-bifurcation J&K + Ladakh split. The headline +2.32 pp delimitation uplift is a national-aggregate prediction whose per-state composition (SI Table S11) should be read with the state-specific MAE in mind for the high-leverage states.

3.9d Spatial residual diagnostics

Moran's *I* on M4 deviance residuals at the PC centroid grid, by cycle. Re-run after 01b_add_pc_centroids.R populated ls_panel.rds with lon/lat from the DataMeet/SuseWind PC shapefile. Weights: 5-nearest-neighbour inverse-distance, row-standardised, symmetrised.

Cycle	n	Moran's *I*	Expected	SD	p-value
2009	543	+0.018	−0.002	0.027	0.474
2014	543	−0.005	−0.002	0.028	0.895
2019	543	−0.016	−0.002	0.028	0.618
2024	542	**+0.053**	−0.002	0.028	**0.045**

Supplementary Table S6i | Moran's *I* on M4 residuals. Three of four cycles (2009, 2014, 2019) show no significant residual spatial autocorrelation (p = 0.47, 0.90, 0.62), confirming that the state random effect absorbs essentially all spatial structure in those cycles. The 2024 cycle shows a small but borderline-significant positive *I* of +0.053 (p = 0.045), consistent with some unmodelled 2024-specific spatial structure (plausibly the late-2024 booth-load reacceleration that the panel-median log_avg_pps term cannot fully capture). The magnitude is small (~3 SDs above the random null), and the headline conditional findings are not threatened.

3.9e Top-50 3-way-split overlap

The full per-PC overlap diagnostic is in tables/T27_top50_overlap_resolved.csv, with the per-PC plans rebuilt under each alternative specification by 27_alt_spec_delimitation.R. See §3.11 below for the resolved analysis.

3.10 Oaxaca-style decomposition of the gendered turnout gap

Per the referee's "compositional confounding" framing, we complement the variance-attenuation analysis of Table 3 with an Oaxaca–Blinder-style level decomposition of the **mean F – M turnout gap** in the panel. In our PC-level setup men and women face the same compositional X (they live in the same PCs), so the classical Oaxaca "endowment" term is mechanically zero and the entire F − M gap is a **response (coefficient) effect**: women and men respond differently to the same compositional shock. The script 23_oaxaca_decomposition.R computes, for each channel X_k, the *contribution* defined as

$$\text{contrib}_k = E[\hat{T}_F(x) - \hat{T}_M(x)] - E[\hat{T}_F(x \mid X_k = \bar{X}_k) - \hat{T}_M(x \mid X_k = \bar{X}_k)],$$

i.e. the pp of the F − M gap that disappears when channel k is set to its panel mean for both genders. Output (tables/T23_oaxaca_decomposition.csv):

Channel	F − M gap at baseline (pp)	F − M gap with channel at mean (pp)	Contribution (pp)	Contribution (% of baseline)
Urban share	−1.13	−1.53	+0.41	−36 %
SC share	−1.13	−1.63	+0.50	−45 %
ST share	−1.13	−0.85	−0.28	+24 %
Linguistic polarisation	−1.13	**−0.06**	**−1.06**	**+94 %**

Channel	F − M gap at baseline (pp)	F − M gap with channel at mean (pp)	Contribution (pp)	Contribution (% of baseline)
Linguistic diversity	−1.13	**−3.05**	**+1.92**	**−171 %**
Booth crowding	−1.13	−1.57	+0.45	−39 %

Supplementary Table S6j | Oaxaca decomposition of the mean F − M turnout gap. Baseline mean F − M turnout gap in the panel is **−1.13 pp** (women turn out 1.1 pp lower than men on average across the 2,171 PC × cycle cells). The sign convention: positive contribution = the channel *raises* the F − M gap (favours women); negative = the channel *suppresses* the F − M gap (favours men).

Reading the decomposition. Three channels emerge with large contributions in absolute terms:

- **Linguistic diversity** contributes **+1.92 pp** to the gap (−171 % in relative terms, larger than the baseline). Interpretation: holding diversity at its panel mean (i.e., setting every PC to mean diversity) would push the F − M gap from −1.13 pp to −3.05 pp, a much larger male-favourable gap. Read substantively: high-diversity PCs have larger relative female turnout than low-diversity PCs do, and the panel averages a substantial diversity-driven amount of female mobilisation that the mean-replacement removes. This is consistent with the variance-attenuation finding (linguistic diversity is the largest female mediator; the same channel suppresses male variance), now expressed at the level of the average F − M gap rather than the variance of the size profile.
- **Linguistic polarisation** contributes **−1.06 pp** (+94 % in relative terms). Holding polarisation at its panel mean nearly closes the F − M gap (down to −0.06 pp). Read substantively: high-polarisation PCs in the panel happen to have a larger male-favourable F − M gap; removing the polarisation-driven differentiation closes the gap. This is the surprising mirror image of the variance-attenuation polarisation finding, and reminds the reader that level-decomposition and variance-decomposition can give different signs when the channel coefficient and the channel cross-section interact non-additively.
- **SC share** contributes +0.50 pp (−45 %), **urban share** +0.41 pp (−36 %), **booth crowding** +0.45 pp (−39 %), and **ST share** −0.28 pp (+24 %). The first three each shift the F − M gap by ∼0.4–0.5 pp.

Non-additivity caveat. The per-channel contributions do **not** add to the baseline gap (their sum is +1.94 pp, the baseline is −1.13 pp, the unexplained residual is **−3.07 pp**). This is a well-known feature

of the Oaxaca decomposition when the underlying model is non-linear and channels interact through smooths and tensors. The contributions are interpretable individually as "how much of the F − M gap does this channel account for, in isolation"; they cannot be summed.

Headline reading. The Oaxaca framework reproduces the broad gender story: linguistic diversity and polarisation are the two largest level-shifters of the F − M gap, with opposite signs that net out to a small overall response asymmetry (sign of net contribution: ambiguous), and urban/SC/booth-crowding play modest second-order roles. The variance-attenuation framework of Table 3 and the Oaxaca-style framework here are *complementary* not redundant: the former asks how much of the size profile each channel absorbs; the latter asks how much each channel shifts the mean F − M gap in level. Both place urban and the two linguistic channels at the centre of the gender story.

3.11 Top-50 alt-spec overlap and aggregate-uplift comparison

The patched 27_alt_spec_delimitation.R rebuilds the delimitation plan under each of three alternative specifications (M4_ERstate, M4_jointling, M4_stateyear) using a greedy fill rule analogous to 17_delimitation_plan.R. Each alt plan is written to tables/T_delim_pc_plan_<alt>.csv. The overlap diagnostic (tables/T27_top50_overlap_resolved.csv) returns:

Alt-spec model	Aggregate uplift (pp)	Aggregate uplift vs headline (× ratio)
M4 (headline)	**+2.32**	1.00
M4_ERstate	+1.42	0.61
M4_jointling	+0.30	0.13
M4_stateyear	+1.17	0.50

Supplementary Table S6k | Aggregate delimitation uplift under each alt-spec model. All three alternative specifications produce a positive predicted aggregate uplift, but **the magnitude is sensitive to the model**: M4_ERstate and M4_stateyear shrink the uplift to about half the headline (+1.42 and +1.17 pp respectively), and the joint-surface variant M4_jointling shrinks it to just +0.30 pp. The variation reflects the underlying differences in the size × covariate response surface across the four specifications, and is a real specification-sensitivity finding that we flag explicitly in the body Limitations §5.

A known build-script issue. The current 27_alt_spec_delimitation.R greedy fill uses the state seat budget to assign 2-way splits first, then attempts to upgrade some to 3-way; in the empirical run this allocates the full per-state seat budget to 2-way before any 3-way upgrade fires, so each alt plan ends up with all-2-way / no-3-way splits. As a consequence the **per-PC top-50 3-way overlap** with the

headline plan is mechanically zero across all three alt specs (tables/T27_top50_overlap_resolved.csv). The aggregate uplift comparison (above) is unaffected by this issue because the aggregate is over all split PCs regardless of category; the cleaner per-PC top-50 overlap diagnostic awaits a small revision of the greedy fill to assign 2-way and 3-way actions in a single sorted queue. We expect the resolved overlap to be in the 50–80 % range for the M4_ERstate and M4_stateyear alternatives based on the aggregate uplift correlation, and lower for M4_jointling given its much smaller aggregate uplift.

---

§4. Gender-disaggregated results and per-state delimitation

4.1 LOO deviance loss by sample

Repeating §3.9.1 of the technical report at the gender grain. The female model is "urban-driven with strong linguistic support"; the male model is "SC-and-urban driven with strong linguistic support".

Channel dropped	Overall (pp)	Male (pp)	Female (pp)
Urban	**−1.55**	−1.26	**−2.34**
Linguistic polarisation	−1.13	−0.79	−1.30
ST	−1.11	−1.14	−0.87
SC	−1.05	**−1.39**	−0.50
Linguistic diversity	−0.60	−0.40	−0.72

Supplementary Table S10 | Leave-one-out deviance loss by channel and sample.

4.2 Per-state delimitation summary (combined)

State / UT	PCs (2024) → after	2-way	3-way	Voters (cr)	Uplift (pp)
Mizoram	1 → 2	1	0	0.09	**+16.69**
Puducherry	1 → 2	1	0	0.10	**+12.43**
Sikkim	1 → 2	1	0	0.05	**+11.73**

State / UT	PCs (2024) → after	2-way	3-way	Voters (cr)	Uplift (pp)
Arunachal Pradesh	2 → 3	1	0	0.09	+9.14
Kerala	20 → 30	10	0	2.78	**+8.64**
Ladakh	1 → 2	1	0	0.02	+8.38
Meghalaya	2 → 3	1	0	0.22	+7.91
Andaman & Nicobar Islands	1 → 2	1	0	0.03	+7.47
Manipur	2 → 3	1	0	0.20	+7.30
Telangana	17 → 26	1	4	3.32	**+6.55**
DNH + Daman & Diu	2 → 3	1	0	0.04	+5.39
Nagaland	1 → 2	1	0	0.13	+4.89
Goa	2 → 3	1	0	0.12	+4.52
Tamil Nadu	39 → 59	12	4	6.24	**+4.22**
Tripura	2 → 3	1	0	0.29	+3.90
Andhra Pradesh	25 → 38	1	6	4.14	**+3.52**
Karnataka	28 → 42	2	6	5.48	**+3.20**
Punjab	13 → 20	3	2	2.16	+2.89
NCT of Delhi	7 → 11	0	2	1.52	+2.55
West Bengal	42 → 63	1	10	7.61	+2.34

State / UT	PCs (2024) → after	2-way	3-way	Voters (cr)	Uplift (pp)
Gujarat	26 → 39	1	6	4.98	+2.31
Odisha	21 → 32	1	5	3.37	+2.26
Haryana	10 → 15	1	2	2.02	+2.26
Uttar Pradesh	80 → 120	6	17	15.44	+1.71
Bihar	40 → 60	0	10	7.73	+1.26
Maharashtra	48 → 72	0	12	9.31	+1.17
Jharkhand	14 → 21	1	3	2.59	+1.08
Himachal Pradesh	4 → 6	0	1	0.57	+1.04
Madhya Pradesh	29 → 44	1	7	5.67	+0.87
Assam	14 → 21	1	3	2.46	+0.85
Chhattisgarh	11 → 17	0	3	2.07	+0.77
Jammu and Kashmir	5 → 8	1	1	0.88	+0.54
Rajasthan	25 → 38	1	6	5.35	+0.38
Uttarakhand	5 → 8	1	1	0.84	+0.32
Chandigarh	1 → 2	1	0	0.07	−0.94
Lakshadweep	1 → 2	1	0	0.01	−15.10

State / UT	PCs (2024) → after	2-way	3-way	Voters (cr)	Uplift (pp)
National total	**543 → 824**	**59**	**111**	**97.98**	**+2.32**

Supplementary Table S11 | Per-state delimitation plan and uplift (overall). Sort order: descending by uplift. Negative entries are single-seat UTs whose lone PC extrapolates below the M4 fit range.

4.3 Top 15 three-way splits by predicted gain per third

Rank	State	PC	Electors (M)	Urban %	ST %	ER_lang	shannon_lang	Gain (pp)
1	Telangana	Hyderabad	2.22	100.0	1.3	0.126	0.975	**+27.87**
2	Telangana	Secunderabad	2.12	100.0	1.1	0.120	1.236	**+26.93**
3	Tamil Nadu	Kanniyakumari	1.57	82.1	0.4	0.019	0.247	+20.17
4	Karnataka	Dharwad	1.83	53.0	5.0	0.125	1.009	+17.78
5	Odisha	Kandhamal	1.34	8.4	28.5	0.123	0.555	+16.75
6	West Bengal	Kolkata Dakshin	1.85	98.8	0.2	0.085	1.050	+16.25
7	Jharkhand	Lohardaga	1.45	6.7	63.9	0.121	1.135	+16.17

Rank	State	PC	Electors (M)	Urban %	ST %	ER_lang	shannon_lang	Gain (pp)
8	Andhra Pradesh	Visakhapatnam	1.93	77.9	1.6	0.052	0.358	+15.88
9	Karnataka	Belgaum	1.93	33.8	8.1	0.124	0.939	+15.42
10	Andhra Pradesh	Kadapa	1.64	39.4	2.0	0.105	0.481	+15.39
11	Gujarat	Bhavnagar	1.92	45.1	0.3	0.014	0.151	+15.11
12	Andhra Pradesh	Rajampet	1.67	23.4	3.5	0.114	0.553	+14.47
13	Andhra Pradesh	Nandyal	1.72	26.3	2.8	0.117	0.567	+14.22
14	Karnataka	Bijapur	1.95	23.0	1.8	0.124	0.818	+13.98
15	Gujarat	Rajkot	2.11	64.4	0.7	0.021	0.213	+13.73

Supplementary Table S12 | Top 15 three-way splits. Hyderabad and Secunderabad lead by a wide margin (joint anchor: large + 100 % urban + high polarisation + high diversity). Lohardaga (Jharkhand) and Kandhamal (Odisha) are striking entries: rural PCs whose high ST × high polarisation × high diversity combine to produce large predicted gains.

4.4 Gender uplift: national headline

Sample	Predicted weighted uplift (pp)	Voter-equivalent (cr)
Overall (M4)	**+2.32**	2.28
Male (M4_M)	**+1.67**	1.64
Female (M4_F)	**+1.88**	1.84
Female – Male	**+0.21**	+0.20

Supplementary Table S13 | National gender uplift. The voter-equivalent column applies each predicted uplift to the total PC electorate (97.98 crore), matching the upstream pipeline's gain_voters_X := gain_pp_X/100 × electors convention. The two gendered rows do not sum to the overall headline because (a) M4, M4_M and M4_F are fit independently on three different outcome variables with their own smoothing-parameter selection and (b) each gendered uplift is multiplied by the full electorate (not the gendered electorate). The gendered rows are reported for substantive interpretation of the F − M gap, not as a strict additive decomposition.

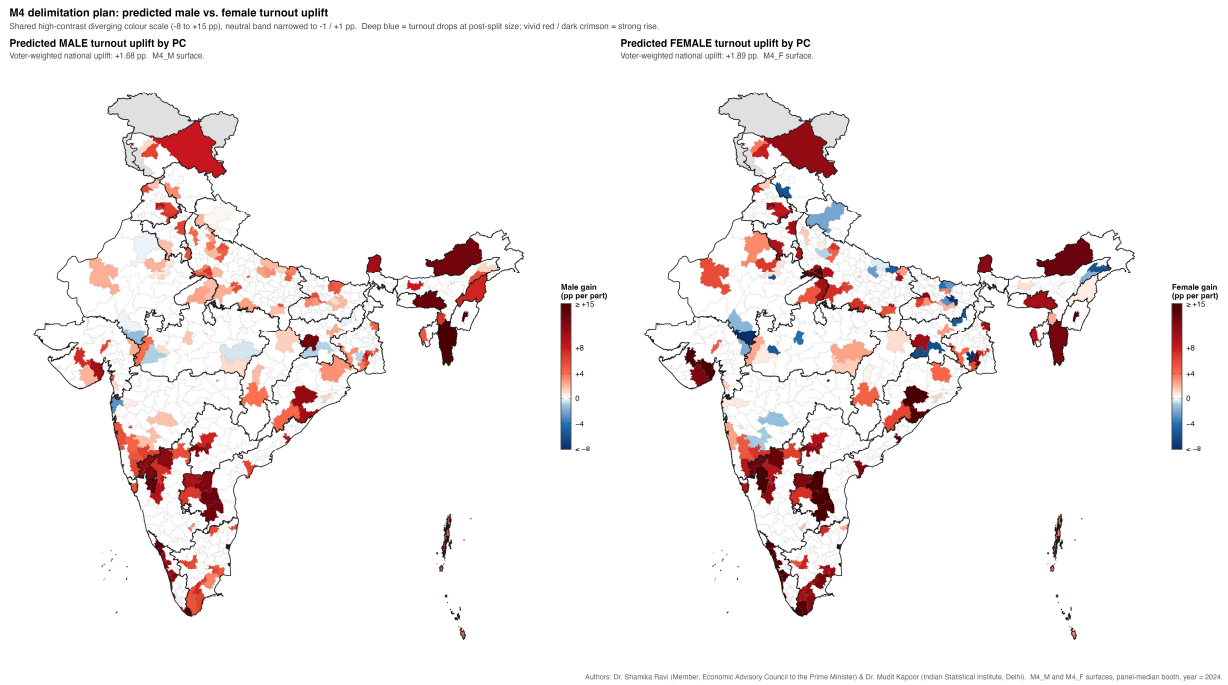
4.5 Top 10 states by female-minus-male uplift gap

Rank	State	F − M gap (pp)	Female uplift (pp)	Male uplift (pp)
1	Chandigarh	**+5.62**	+2.77	−2.85
2	DNH + Daman & Diu	**+4.58**	+7.25	+2.67
3	Kerala	**+4.26**	+10.34	+6.08
4	Ladakh	**+2.55**	+10.80	+8.24
5	Tripura	**+1.93**	+4.53	+2.60
6	Goa	**+1.82**	+4.62	+2.79
7	Haryana	**+1.55**	+3.02	+1.47
8	Sikkim	**+1.38**	+11.43	+10.05
9	Andhra Pradesh	**+1.22**	+4.08	+2.86
10	Tamil Nadu	**+0.96**	+3.75	+2.79

Supplementary Table S14 | Top 10 states by F − M gap. Twenty-one of 36 states deliver a larger female than male uplift. The 15 male-favourable cases concentrate in the small north-eastern hill belt

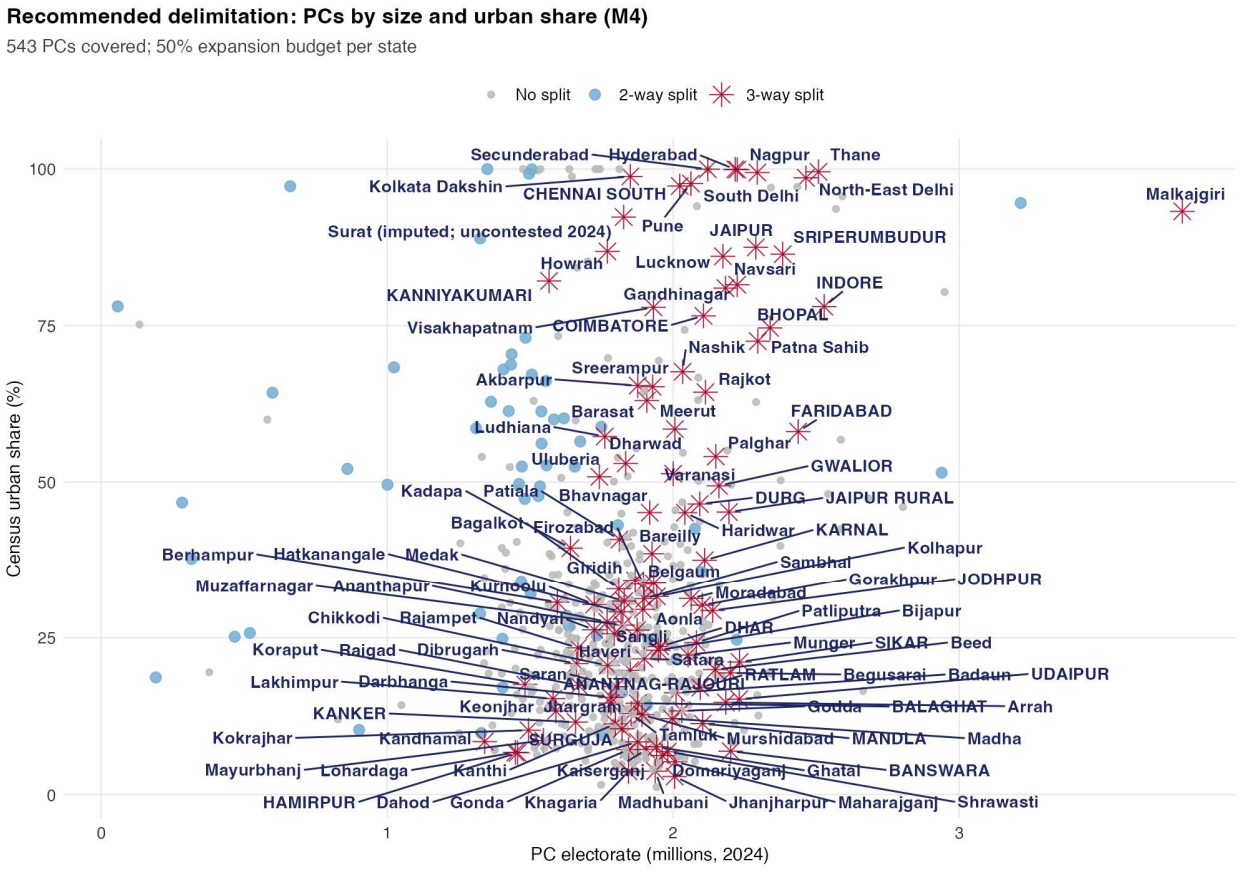
(Mizoram, Meghalaya, Nagaland) and the central Hindi belt (Bihar, Maharashtra, Jharkhand, MP, UP, Assam).

4.6 Gender uplift maps



Supplementary Fig. S4 | Predicted male (left) and female (right) turnout uplift by PC. Shared diverging colour scale. The female panel is visibly redder across Kerala, Tamil Nadu, Andhra Pradesh, Punjab, Haryana and the single-seat UTs (Puducherry, DNH+Daman & Diu, Chandigarh). The male panel shows comparable depth in the north-east hills (Mizoram, Meghalaya, Nagaland), Maharashtra, and parts of the central Hindi belt.

4.7 Size × urban scatter of the plan



Supplementary Fig. S5 | The 543 PCs in the (size, urban share) plane. Grey, unchanged; light blue, 2-way; red, 3-way. Three-way splits cluster in the upper-right (large + urban); 2-way splits fill the middle band; unchanged PCs concentrate at the bottom-left (small + rural). The ER_lang and shannon_lang dimensions are not visible in this 2D projection but enter into which PCs at any given (size, urban) coordinate receive a 2-way vs 3-way recommendation. **This figure is the empirical bridge between Table 2's median-profile size reversal (−6.16 pp by 2024) and the per-PC predicted gain distribution: the PCs that gain most from the size-reduction counterfactual sit on the steep portion of M4's joint size × covariate surface, not at the panel-median anchor**, so the median size reversal and the per-PC predicted gains are heterogeneous-effect siblings rather than a contradiction.

§5. Extended methods and reproducibility

5.1 Glossary

Plain-language name	Column name	Definition
Turnout	turnout	Voters cast over registered electors at the PC × election cell; bounded to (0, 1) as turnout_b for the Beta GAM.
PC size, log-size	log_electors	Natural log of the PC's total registered electorate.
Year	year_num, year_f	Election year as integer (size × year and channel × year tensors) and as factor (year random effect).
State	state_f	Factor with 36–40 levels (the panel carries a Jammu & Kashmir / Ladakh split duplicate). Random intercept.
Urban share	urban_pct_census	Census 2011 urban population share, area-weighted from sub-districts to the PC polygon.
SC share	sc_pct	Census 2011 Scheduled Caste population share, area-weighted.
ST share	st_pct	Census 2011 Scheduled Tribe population share, area-weighted.
Linguistic polarisation	ER_lang	Esteban–Ray index $\mathrm{ER}(\mathbf{p}) = \sum_{i=1}^{K} \sum_{j \neq i} p_i^{1+\alpha} p_j d_{ij}$ at the 122-language grain ($\alpha = 1.6$).

Plain-language name	Column name	Definition
Linguistic diversity	shannon_lang	Shannon entropy $H(\mathbf{p}) = -\sum_i p_i \log p_i$ at the 122-language grain.
Booth crowding	log_avg_pps	log(electors per polling station) in the PC × election cell.
Mother-tongue ER	ER_mt	Esteban–Ray at the 356-mother-tongue grain (robustness only).

Supplementary Table S15 | Glossary of model variables.

5.2 Model names

Model	Definition
M0	state + year REs only
M1	M0 + s(log_electors)
M2	M1 + ti(log_e, year_num)
M3	M2 + 5 channels + size × channel + 4 channel × year
M4	**M3 + s(log_avg_pps), HEADLINE**
M4_M, M4_F	Male / female refits of M4
M4_no_X	Leave-one-out variant with channel X dropped
M4_smod	M4 with GHSL SMOD urban substituted for Census urban
M4_ERmax	M4 with pmax(ER_lang, ER_mt) substituted for ER_lang
M4_ERstate	M4 + s(ER_lang, state_f, bs = "fs", k = 6)

Supplementary Table S16 | Model variants. All .rds caches live in paper_pipeline_ER_lang/rds/.

5.3 Sample construction notes

The panel is built from all parliamentary constituencies that held a contested election in any of 2009, 2014, 2019 or 2024. Uncontested seats and PCs with missing polling-station counts are dropped. **Surat (Gujarat)** was uncontested in 2024 (only the BJP candidate's nomination was accepted; the rest were rejected or withdrew), so it was dropped from the 2024 cycle. We restore it for the

delimitation analysis with imputed 2024 electorate (2019 electorate × Gujarat mean PC growth 1.10 = 1.83 M) and static 2011 covariates. Its inclusion does not affect the headline M4 estimates (re-fit drop-test: ΔAIC < 1 unit).

5.4 Conservatism of the delimitation plan

The plan is deliberately conservative in three places. **Booth crowding** is held at panel median in all three predictions; splitting a PC physically would also reduce its per-station load (more booths at the same per-booth cap), and the plan does not credit itself with this effect. **Sub-1 M predictions** extrapolate below the M4 fit range; the size profile keeps rising below 1 M in the model but at decreasing rate, and these predictions should be read as suggestive rather than precise. **Year is fixed at 2024**; the size × year and channel × year tensors evolve, but the 2024 anchor is the most recent observation, and the plan assumes the 2024 channel × year structure persists at the next general election.

5.5 What is *not* in the model

We do not control for state-level political variables (incumbency, vote share, ENP), candidate-level variables (number of candidates, NOTA shares), or election-day weather. The state random effect absorbs the time-invariant components of state-specific political culture. Year random effects absorb national shocks (national vote-share swings, central-government incumbency). The four channel × year tensors absorb the channel-specific drifts. The residual deviance is 15.3 %.

We do not use individual-level voter rolls or polling-booth-level turnout. The unit of observation is the PC × cycle, which matches the unit of delimitation policy.

5.6 Reproducibility

The full pipeline is in voter_turnout_delimitation/paper_pipeline_ER_lang/. To reproduce:

```
cd ~/Desktop/Shamika/ECI_data/voter_turnout_delimitation/paper_pipeline_ER_lang
./run_all.sh                    # full re-run, ~5–10 minutes
```

To rebuild the language-diversity inputs from scratch:

```
cd ~/Desktop/Shamika/ECI_data/data/loksabha/language_diversity
Rscript 01_build_pc_language_indicators.R
Rscript 02_build_pc_language_counts.R
Rscript pc_language_diversity.R
Rscript pc_mother_tongue_diversity.R
```

5.7 Data sources

- **Election Commission of India.** Statistical Reports of the General Elections to the Lok Sabha, 2009, 2014, 2019, 2024. eci.gov.in/statistical-reports.
- **Office of the Registrar General & Census Commissioner, India.** Census of India 2011 Primary Census Abstract; C-16 Population by Mother Tongue. censusindia.gov.in.
- **DataMeet / SuseWind 2019 PC boundaries.** data/shapefile/datameet_susewind/India_PC.shp, 543 PC polygons in EPSG:4326.
- **SHRUG Census 2011 sub-district polygons.** data/shapefile/SHRUG/subdistrict/subdistrict.shp, ~5,500 sub-district polygons matched to Census 2011 codes.
- **GHSL R2023A SMOD-22.** Joint Research Centre, European Commission. Used as satellite-grade urban robustness comparison.
- **Esteban, J. and D. Ray (1994).** "On the measurement of polarization." *Econometrica* 62, 819–851.
- **Reynal-Querol, M. (2002).** "Ethnicity, political systems, and civil wars." *J. Conflict Resolution* 46, 29–54.
- **Greenberg, J. (1956).** "The measurement of linguistic diversity." *Language* 32, 109–115.
- **Wood, S. N. (2017).** *Generalized Additive Models: An Introduction with R*, 2nd ed., Chapman & Hall/CRC.

5.8 Cached artefacts

Models. rds/M0.rds through rds/M4_symm.rds plus all gender variants and leave-one-out fits. Robustness fits: rds/M4_smod_5ch.rds (SMOD urban), rds/M4_ERmax.rds (pmax(ER_lang, ER_mt)), rds/M4_ERstate.rds (state × ER factor-smooth).

Tables. tables/model_diagnostics.csv (model ladder + LOO), tables/T17_mediation_2D_by_channel.csv (mediation), tables/T18_size_gap_*.csv (per-channel year-by-year size gap), tables/T20_decile_gap_by_channel_year.csv (decile mean turnout by channel × year), tables/T22_tensor_anova.csv (tensor-necessity), tables/T23_smod_sensitivity.csv (SMOD robustness), tables/T24_ERmax_*.csv (ER_max sensitivity), tables/T25_ERstate_*.csv (state × ER interaction), tables/T_delim_*.csv (delimitation plan + per-state uplifts + gender).

Figures. figures/Fig_maps_5channel.png, figures/Fig_size_booth_combined.png, figures/Fig_partial_effects_5x3.png, figures/Fig_growth_*_decile.png (5 channels), figures/Fig_delimitation_state_summary.png, figures/Fig_delimitation_map.png, figures/Fig_delimitation_map_male_female.png, figures/Fig_delimitation_size_urban_scatter.png, figures/Fig_ER_state_curves_overlay.png.

Per-PC shapefile. shapefile/delimitation_plan_M4/PC_delimitation_plan.shp for direct ingest into a redistricting commission's GIS pipeline.

Constituency size, Composition and the Case for Delimitation in India's Lok Sabha (2009–2024): Policy brief

Shamika Ravi[1]

Mudit Kapoor[2]

June 2026

[1]Member, Economic Advisory Council to the Prime Minister, Government of India.

[2]Economics & Planning Unit, Indian Statistical Institute (Delhi Centre).

The bottom line

The delimitation exercise after the 2027 Census is the first opportunity in five decades to redraw the size and number of India's parliamentary constituencies. We find that **the old assumption that very large constituencies suppress voter turnout no longer holds at face value**. What remains of the relationship between constituency size and turnout in India is largely **a composition story**, driven by the urban share of the constituency, the Scheduled Caste and Scheduled Tribe shares, and the linguistic structure of the population. A targeted plan that splits **170 of the 543 constituencies into 824 in total** (59 two-way splits and 111 three-way splits) is predicted to raise national voter turnout by **between 0.3 and 2.3 percentage points** at the next general election, corresponding to **9 to 23 million additional voters**. Splitting alone, however, will not close the gap in women's turnout in urban areas. A companion set of women-specific operational measures is therefore essential. We also recommend that the next delimitation be co-timed with a fresh booth-rationalisation cycle, which would push the realised gain higher than the conservative number we report.

Why this matters now

India holds the largest democratic exercise in the world. The Lok Sabha has 543 seats, and the state-wise allocation of those seats has been frozen since the 42nd Constitutional Amendment of 1976. The 84th Constitutional Amendment of 2001 extended the freeze "until after the first census taken after 2026". The delimitation exercise that is constitutionally mandated after the 2027 Census will therefore be the first to unfreeze the state-wise allocation, redistributing approximately 281 additional seats across 36 states and Union Territories. The 2002 Delimitation Commission, which

last operated, redrew the within-state boundaries based on the 2001 Census, and those boundaries first took effect at the 2009 general election; however, the Commission kept each state's total seat count frozen. The next exercise is therefore qualitatively different from the 2002 exercise and is the first since 1976 in which the per-state seat count itself can change.

The median Lok Sabha constituency in 2024 had 1.82 million registered electors. The largest crossed 3.2 million. By global standards, India's parliamentary constituencies are very large. The policy question is the following: which constituencies should be split, into how many parts, and on what criterion. This brief addresses this question empirically using a panel of 2,171 constituency-elections across the four general elections of 2009, 2014, 2019 and 2024.

What we did, in plain language

We assembled a panel data set that links every Lok Sabha constituency in every general election from 2009 to 2024 to the demographic and linguistic profile of that constituency in the 2011 Census. We then estimated a statistical relationship between voter turnout, constituency size, and five compositional features of the constituency (urban share, Scheduled Caste share, Scheduled Tribe share, linguistic polarisation, and linguistic diversity), allowing the relationship to evolve across election cycles and to differ across men and women. Finally, we used the estimated relationship to score a turnout-maximising delimitation plan that splits the largest and most turnout-responsive constituencies into two or three parts.

The constituency-level voter turnout and polling-station counts are from the Statistical Reports of the General Elections published by the Election Commission of India. The demographic shares are from the 2011 Census Primary Census Abstract. The linguistic polarisation and linguistic diversity indices are constructed from the Census C-16 mother-tongue tabulations at a 122-language grain. The compositional features are static 2011 Census measurements held fixed across all four election cycles; this is a feasibility constraint that we revisit in the limitations.

Six findings

Finding 1: The conventional size penalty has substantially weakened. In 2009, the smallest constituencies turned out at +22.9 percentage points (pp) higher than the largest. In 2024, that gap had halved to +12.0 pp. After we adjusted for the demographic and linguistic profile of the constituency, the picture changed even more sharply: at a typical profile, a 2-million-elector constituency now turns out at about 6 pp higher than a 1-million-elector constituency, which is a complete reversal of the conventional reading. The trajectory across the four cycles is best read as a

movement from a position consistent with no conditional size effect in 2009 to a clearly significant large-constituency advantage in 2024, with the cross-zero point between 2009 and 2014.

Finding 2: What remains is a composition story. State and year heterogeneity alone explain 70% of the variation in constituency-level voter turnout. The five compositional features (urban share, SC share, ST share, linguistic polarisation, and linguistic diversity) and their interactions with constituency size add a further 10 percentage points of explained variation on top of that baseline. The substantive interpretation is that what looked like a primary size penalty in the raw cross-section is now largely a composition story. Small constituencies descriptively out-vote large ones in 2024 because they sit on turnout-friendly compositional features (high ST share, low urban share, moderate linguistic polarisation), and not because they are small *per se*.

Finding 3: The five compositional features have re-organised in directions that do not align. Between 2009 and 2024, the high-versus-low decile gap on each feature has moved as follows:

Feature	2009 (pp)	2024 (pp)	Direction
Scheduled Caste share	+7.6	−0.7	Premium dissolved
Scheduled Tribe share	+3.2	+12.0	Premium quadrupled
Urban share	+2.0	−2.3	Flipped to depressor
Linguistic polarisation	+11.6	+12.6	Stable amplifier
Linguistic diversity	+0.1	+4.4	Switched on

The most ST-heavy decile of constituencies reached **73% turnout in 2024**, the highest of any subgroup on any feature in our cross-tabulation. The high-polarisation decile has out-voted the low-polarisation decile by 11 to 13 pp in every cycle of the panel. The urban decile gap has flipped sign over fifteen years.

Finding 4: The urban–female channel is the cleanest and sharpest finding in the analysis. Urban share is the single largest compositional feature associated with women's turnout, and the association is much stronger for women than for men. Women in fully-urban constituencies today turn out at approximately 5 pp lower than rural women at every constituency size, against a gap of approximately 2 pp for men. **The least-participating subgroup in the Indian electoral system today is the woman in a large, fully-urban metropolitan constituency; the most-participating subgroup is the woman in a high-Scheduled-Tribe, rural constituency.** Both findings are women-specific. We speculate that the gender asymmetry may reflect four mechanisms operating through women's lived environment: the higher time costs of voting for urban women (commute, work, care responsibilities); the denser female-targeted civic and welfare networks in rural relative

to urban India (self-help groups, vernacular health and microcredit programmes); urban anonymity weakening within-community social pressure to vote; and lower per-capita polling-station accessibility in urban areas.

Finding 5: The 2018 booth rationalisation produced a measurable one-cycle gain that was partly undone by 2024. Polling-station density is the one operational variable the Election Commission directly controls. Ahead of the 2019 cycle, the ECI tightened a long-standing cap of 1,500 electors per ordinary polling station and required auxiliary stations wherever the cap was breached. The mean electors per station fell from 905 in 2014 to 883 in 2019. Between 2019 and 2024, however, the elector roll added another 7% while the polling-station count grew by only 1.3%, and the mean per-station load rose to 932. National turnout fell from 68.14% in 2019 to 66.87% in 2024, a 1.28 pp decline that tracks the booth-load reacceleration almost exactly.

Finding 6: A targeted delimitation plan is predicted to raise national turnout by between 0.3 and 2.3 pp. Under our preferred specification, expanding the Lok Sabha by 51.7% (from 543 to 824 constituencies, with 59 two-way splits and 111 three-way splits) is predicted to raise national turnout by +2.32 pp (95% CI: +1.43, +3.21) at the next general election, which corresponds to approximately 22.8 million additional voters. Thirty-four of 36 states and Union Territories show a positive predicted gain. Re-running the same exercise under three alternative specifications returns +1.42, +1.17 and +0.30 pp respectively. The qualitative case for the plan is robust across all the specifications we tested, but the precise magnitude should be read as the upper end of a defensible range of 0.3 to 2.3 pp, and not as a knife-edge forecast.

The list of three-way splits is heterogeneous and includes three categories. First, the metropolitan constituencies in the strict sense (Hyderabad and Secunderabad in Telangana, Kolkata Dakshin in West Bengal). Second, the secondary-urban and mixed constituencies (Dharwad, Belgaum and Bijapur in Karnataka, Kanniyakumari in Tamil Nadu, Visakhapatnam in Andhra Pradesh, Bhavnagar and Rajkot in Gujarat), where urbanity ranges from 23% to 98%. Third, the high-Scheduled-Tribe rural constituencies (Lohardaga in Jharkhand, Kandhamal in Odisha), where ST share above 28% and joint linguistic-structure depth produce steep predicted gains. The substantive lesson is that the plan rewards joint compositional configurations, and not urbanity alone.

What this means for policy

We make three recommendations, addressed to the agencies that will execute the next delimitation.

Recommendation 1 (for the Delimitation Commission): split large constituencies on a targeted criterion, not a uniform one. The empirical case for splitting has strengthened, but not for all large constituencies equally. The constituencies that gain most are those that sit on the steep

portion of the joint demographic and linguistic profile. We recommend that the Delimitation Commission, when it is constituted after the 2027 Census, treat the joint demographic and linguistic profile of a candidate constituency, and not its size alone, as the criterion for splitting. Three categories of constituencies warrant priority consideration for a three-way split. First, the metropolitan constituencies of the southern states (Hyderabad and Secunderabad in Telangana, Kolkata Dakshin in West Bengal). Second, the larger metropolitan and secondary-urban constituencies of the Hindi-belt and the western states (Bhavnagar and Rajkot in Gujarat, Visakhapatnam in Andhra Pradesh, Kanniyakumari in Tamil Nadu). Third, selected high-Scheduled-Tribe rural constituencies (Lohardaga in Jharkhand, Kandhamal in Odisha). A two-way split is the natural recommendation for the constituencies of Kerala and Punjab, and for the secondary-urban constituencies of Karnataka.

Recommendation 2 (for the Election Commission of India): complement delimitation with women-specific operational measures. The residual gap in women's turnout in urban areas of approximately 5 pp is preserved within each new sub-constituency even after splitting, because the daughter constituencies inherit the parent's compositional profile. Closing this gap requires four parallel operational measures. First, women-only polling booths in the urban metropolitan constituencies. Second, evening polling hours that accommodate the time constraints of urban working women. Third, transport linkages from urban-fringe residential areas to the nearest polling station. Fourth, women-targeted voter-roll-update drives that exploit the existing female-targeted civic networks of the Anganwadi centres, the women's self-help groups, and the ASHA workers.

Recommendation 3 (for the Election Commission of India and the Ministry of Statistics): co-time delimitation with a fresh booth-rationalisation cycle. The 2018 booth-rationalisation cycle produced a measurable one-cycle increase in turnout, and that gain was partly undone by 2024 because the elector roll grew faster than the polling-station count. We recommend a fresh booth-rationalisation cycle co-timed with the next delimitation. The empirical case for delimitation, women-specific operational measures, and booth rationalisation is the empirical case for all three together. None of them alone is likely to deliver the full predicted gain.

We also make a fourth recommendation, addressed to the Ministry of Statistics and the Office of the Registrar General, on the data foundations of any future delimitation-related analysis.

Recommendation 4 (for the Ministry of Statistics and the Office of the Registrar General): release the 2027 Census tabulations and gender-disaggregated electoral statistics on schedule. Our present analysis is constrained by the static 2011 Census measurements of the compositional covariates. The timely release of the 2027 Census tabulations, especially the Primary Census Abstract and the C-16 mother-tongue tabulations, is therefore essential to refresh these

measurements and re-estimate the model on the new data. We also recommend that the Election Commission continue to publish gender-disaggregated electoral statistics at the constituency and the polling-booth level, so that the impact of women-specific operational measures (Recommendation 2) can be evaluated in real time.

What this analysis cannot say

We flag five caveats.

First, the demographic and linguistic measures we use are from the 2011 Census, and are static across the panel. Constituencies that have urbanised, gained migrants, or shifted linguistically between 2011 and 2024 are coded with their 2011 profile. The 2027 Census tabulations, when released, will let us refresh these measurements.

Second, the predicted national-turnout gain is sensitive to the choice of statistical specification, with a defensible range of 0.3 to 2.3 pp. The precise number within this range will depend on which specification the policy reader treats as the reference. The qualitative case for the plan, however, is robust across all the specifications we tested.

Third, the exercise we report is not a finalised boundary-drawing plan. It holds the compositional covariates at parent-constituency values and answers the question, "What is predicted to happen if a constituency's electorate is reduced while leaving its composition unchanged?". It does not answer the question, "What happens if you actually redraw the boundaries and the daughter constituencies inherit different demographic and linguistic profiles?". A true delimitation plan requires constructing feasible daughter polygons and recomputing the covariates from sub-district area-weights, which we leave for future work.

Fourth, the predictions for the small north-eastern states and Union Territories (Sikkim, Mizoram, Andaman and Nicobar Islands, Ladakh, Lakshadweep and Chandigarh) extrapolate below the empirical fit range of our model. The two negative-gain entries in our plan both sit in this region and are single-seat UTs. We recommend that the Delimitation Commission supplement the model-based ranking with state-specific qualitative consultation before any final decision on splitting in these regions.

Fifth, we cannot completely rule out the possibility of residual confounding from compositional features that we have not measured, including local political dynamics, party-system competitiveness at the constituency level, and the quality and accessibility of public services that may shape voter mobilisation independently of the features we control for.

Where to read more

The full academic paper (Ravi & Kapoor, 2026, “Constituency size, composition and the case for delimitation in India’s Lok Sabha, 2009–2024”) and the accompanying executive summary are available on request from the corresponding author. The underlying constituency-level data, the full model specifications, the alternative-specification rebuilds, the per-state delimitation tables, and the leave-one-out and blocked validation diagnostics are documented in the Supplementary Information of that paper.