NYC 2025 Mayoral Election Forecast Model: Two-Compartment Multinomial Ridge with Offsets

Peter Driscoll

November 4, 2025

Purpose

This document is the authoritative specification for the NYC 2025 mayoral forecast. It defines the target, model form, priors, features, turnout assumptions, estimation, calibration, and outputs. It is the single source of truth for implementation and communication.

1 Objective and Unit of Analysis

Target: Election District (ED)–level *vote shares* for three candidates $\{k\} = \{\text{Mamdani}, \text{Cuomo}, \text{Sliwa}\}$, aggregatable to AD/borough/city. We forecast *shares* (not totals). Turnout is used to shape composition and aggregation weights, not to directly predict counts.

2 Model Overview

We model two disjoint "compartments" per ED i: EV (banked early voters) and R (remaining/Election Day voters). Within each compartment we use a multinomial logistic model with an offset prior, then mix the two by their ED-specific turnout weights.

2.1 Compartment models (per candidate k)

$$\eta_{ik}^{(c)} = o_{ik} \; + \; \gamma_k^{(c)} \; + \; \beta_k^{\top} \mathbf{x}_i^{(c)}, \qquad c \in \{\text{EV}, \text{R}\}, \tag{1}$$

$$p_{ik}^{(c)} = \frac{\exp\left(\eta_{ik}^{(c)}\right)}{\sum_{\ell} \exp\left(\eta_{i\ell}^{(c)}\right)}, \qquad \sum_{k} p_{ik}^{(c)} = 1. \tag{2}$$

2.2 Two-compartment mixture to ED-level share

Let T_i be the projected total turnout for ED i under a turnout scenario (Sec. 4). Define weights

$$w_i^{\text{EV}} = \frac{\text{EV}_{25,i}}{T_i}, \qquad w_i^{\text{R}} = 1 - w_i^{\text{EV}}.$$

The predicted ED share is

$$p_{ik} = w_i^{\mathrm{EV}}\, p_{ik}^{(\mathrm{EV})} \;+\; w_i^{\mathrm{R}}\, p_{ik}^{(\mathrm{R})}, \qquad \sum_k p_{ik} = 1. \tag{3}$$

3 Offsets (Pattern Priors)

Offsets o_{ik} anchor spatial patterns using the freshest and most relevant signals.

- **Mamdani:** $o_{iM} = \mathrm{logit}\,\big(\text{Dem-primary Mamdani share in ED } i\big)$.
- **Cuomo:** $o_{iC} = 0$ (no direct Democratic-primary analog; learns from features and calibration).
- **Sliwa (weak stabilizer):** $o_{iS} = \alpha_S\ \mathrm{logit}\,\big(\text{Sliwa 2021 share in ED } i\big)$ with small α_S to avoid over-weighting history while providing a plausible GOP spatial pattern.

All non-offset effects are shrunk via ridge (L2) to promote stability with small, noisy tabular data.

4 Turnout Scenarios and Weights

We do not directly model totals. Instead, turnout shapes composition and aggregation via T_i and the compartment weights.

- **Total turnout scenarios:** $T \in \{1.80\text{M}, 1.94\text{M}\}$.
- **Cannibalization of E-Day by EV:** $\rho \in \{0.4, 0.6\}$ is the fraction of 2021 Election Day volume cannibalized by 2025 EV.
- **Per-ED projection:** Let $\mathrm{EDay21}_i = \mathrm{Total21}_i - \mathrm{EV}_{21,i}$. Define

$$\widehat{\mathrm{EDay25}}_i \;=\; (1-\rho)\, s \cdot \mathrm{EDay21}_i,$$

  where s scales so that $\sum_i \big(\mathrm{EV}_{25,i} + \widehat{\mathrm{EDay25}}_i\big) = T$. Then $T_i = \mathrm{EV}_{25,i} + \widehat{\mathrm{EDay25}}_i$ and $w_i^{\mathrm{EV}} = \mathrm{EV}_{25,i}/T_i,\ w_i^{\mathrm{R}} = 1 - w_i^{\mathrm{EV}}$.

5 Feature Sets (MECE by compartment)

All features are standardized (z-scores). Use a compact set to avoid collinearity; ridge will shrink noisy terms.

EV compartment (banked voters): $\mathbf{x}_i^{(\mathbf{EV})}$

- $\mathrm{youth\_prop}_i$: ACS share age 18–29.
- $\mathrm{student\_share}_i$ *or* $\mathrm{near\_campus}_i$ (choose one).
- $\%\mathrm{EV\_newly\_registered}_i$ (best) *or* $\mathrm{new\_reg\_rate}_i = \frac{\text{new regs 2024–25}}{\text{registrants}}$.
- $\mathrm{income\_above\_median}_i$ (or tract income percentile).
- Optional access: $\mathrm{ev\_site\_near}_i$ or distance.

Remaining / E-Day compartment: $\mathbf{x}_i^{(\mathbf{R})}$

- income_below_median$_i$ (complement to EV).
- remaining_gap$_i$ = target_bin_EV_rate − ED_EV_rate$_i$ (who has not voted yet).
- renter_share$_i$ *or* recent_mover_share$_i$ (pick one).
- Retain youth_prop$_i$ (late youth may vote on E-Day).

Candidate-specific stabilizer

- **Sliwa only:** gop_base$_i$ = Sliwa2021_share$_i$ with strong shrinkage (tiny ridge effect).

6 Estimation and Regularization

We fit ridge-penalized multinomial logits in each compartment (or a single model with EV/E-Day interactions if preferred operationally). Choose the ridge parameter λ via *spatial cross-validation* (folds grouped by AD or borough) to prevent leakage from neighboring EDs.

7 Poll Anchoring (Citywide Calibration)

Let $\bar{p}_k$ be the turnout-weighted citywide mean:

$$\bar{p}_k \;=\; \frac{\sum_i T_i\, p_{ik}}{\sum_i T_i}.$$

We apply small intercept shifts to $\gamma_k^{(\mathrm{EV})}$ and/or $\gamma_k^{(\mathrm{R})}$ (or a shared γ_k) so that $\bar{p}_k$ matches the latest polling averages.
Soft Sliwa prior (count-to-share): historical $\approx$ 316k votes implies $\sim$ 16%−18% share when $T \in [1.80, 1.94]$M; we encode this as a weak citywide penalty on $\bar{p}_S$ (polls remain the primary anchor).

8 Uncertainty and Simulation

We quantify uncertainty via:

- Bootstrap over EDs (resampling rows).
- A shared citywide shock added to linear predictors to induce correlation across EDs.

We run simulations across the grid $(T, \rho) \in \{1.80, 1.94\} \times \{0.4, 0.6\}$ and report 50% and 90% intervals for ED/AD/borough/city aggregates.

9 Data Inputs (Outstanding)

1. Democratic primary by ED (votes and shares).
2. EV 2025 by site/day (allocated to ED) and EV 2021.

3. Registration: current totals; new registrations 2024–25; party mix if available.
4. ACS: age 18–29; student share or near-campus flag; income variable.
5. Sliwa 2021 share by ED (or GOP registration share).
6. Numeric polling averages for Mamdani, Cuomo, Sliwa (used in calibration).

10 Implementation Plan (Order of Operations)

1. Build the ED design matrix: offsets o_{ik} and features $\mathbf{x}_i^{(\mathrm{EV})}, \mathbf{x}_i^{(\mathrm{R})}$; standardize features.
2. Compute $T_i, w_i^{\mathrm{EV}}, w_i^{\mathrm{R}}$ for each (T, ρ) scenario.
3. Fit compartment models with ridge; select λ via spatial CV.
4. Poll-align intercepts using turnout-weighted citywide means; include weak Sliwa share prior.
5. Simulate uncertainty across (T, ρ); export ED/AD/borough/city shares with 50/90% bands; render maps and a short methods note.

11 Summary

This specification combines (i) fresh *pattern priors* (Dem primary and a weak GOP map), (ii) composition-aware compartment weights derived from observed early voting and conservative E-Day scaling, and (iii) citywide calibration to current polling. The result is an interpretable, stable forecast of *vote shares* that updates coherently as early-vote and registration information evolves.

Update: Data Integration and Next Steps (November 2025)

Summary of Completed Work

Since the original model specification, we have now constructed and validated the empirical foundation required for estimation.

Data Foundations.

- **Primary data:** Built `primary_final_round_by_ed.csv` from full 2025 Democratic primary cast-vote records. Each Election District (ED) now has:
  - Final-round (top-two) votes for Mamdani and Cuomo.
  - Total valid ballots and computed ED-level shares.
  - Borough inference from Assembly Districts, verified against official borough totals (*match within 0.5pp*).
- **Early voting data:** Extracted from the interactive *THE CITY* dataset and cleaned into `early_votes_by_ed.csv`, containing:
  - Early vote totals by ED, along with party composition (DEM, REP, OTH).

  - Normalized rates (`early_vote_rate`, `early_vote_weight`) and borough/neighborhood mapping.

- **Merged model inputs:** Produced `model_inputs_by_ed.csv` by joining the primary and early vote files on (`AD, ED`). The file now contains:
  - Validated Mamdani/Cuomo shares and logit transforms.
  - Turnout priors (`early_vote_rate`, `early_vote_weight`).
  - Full dtype enforcement and borough backfill via AD inference.

Validation and Diagnostics.

- A dedicated `validate_merge.py` script confirms schema, type consistency, and coverage (100% of EDs matched).
- Borough totals exactly reproduce official final-round results (Cuomo leads Bronx & Staten Island; Mamdani leads Brooklyn, Queens, Manhattan).
- Visual diagnostics confirm expected correlations:
  - $r(\text{early vote rate}, \text{Mamdani share}) \approx +0.20$ (weak positive).
  - Continuous, realistic ED-level distributions of turnout and support.

Outstanding Integrations

- **Sliwa 2021 baseline:** add ED-level GOP baseline as the stabilizing offset for the Republican compartment.
- **Registration data:** integrate borough/ED-level registration totals and new registrations (2024–25). These will anchor turnout weights and potential youth/new-voter priors.
- **ACS / demographic features:** append standardized age, income, and renter/student variables for both compartments. Current early-vote data already provide turnout composition but not demographic mix.
- **Polling anchor:** ingest latest citywide polling averages to calibrate compartment intercepts.

Next Steps

1. **Finalize design matrix:** construct $x_i^{(EV)}$ and $x_i^{(R)}$ feature blocks using merged ED data plus ACS/demographic supplements.
2. **Turnout weights:** compute T_i, w_i^{EV}, and w_i^R under scenarios $(T, \rho) \in \{1.80, 1.94\} \times \{0.4, 0.6\}$, using EV totals as fixed.
3. **Fit compartment models:** estimate ridge-regularized multinomial logits per compartment; cross-validate λ by borough grouping.
4. **Poll calibration:** align intercepts to polling means; apply soft Sliwa prior.

5. **Simulation layer:** run uncertainty draws across turnout and cannibalization scenarios; aggregate to AD, borough, and citywide results with 50/90% credible intervals.
6. **Scenario priors (optional):** use global age/new-registration statistics only as calibration checks or global intercept adjustments (do not use as ED-level features until spatial variation is available).

Deliverables (in progress)

- `primary_final_round_by_ed.csv` — verified final-round primary results.
- `early_votes_by_ed.csv` — normalized 2025 early vote dataset.
- `model_inputs_by_ed.csv` — unified modeling base.
- `validate_merge.py` — reproducible schema/coverage audit.
- `build_final_round_by_ed.py` — robust source-builder with automatic finalist verification.
- Next: `fit_compartments.R` or `fit_ridge.py` for estimation and poll anchoring.

Status Summary

All major data inputs for the Democratic primary and early vote compartments are complete, validated, and ready for feature engineering. The remaining work focuses on:

- enriching demographic covariates,
- encoding turnout scenario logic,
- fitting and calibrating the two-compartment ridge model.

This version supersedes Section 9 (Data Inputs) and Section 10 (Implementation Plan) of the original specification.