PoliSim

Piattaforma di Simulazione Elettorale

White Paper Tecnico — Metodologia e Validazione

Versione 1.0 — Maggio 2026

polisim.dev

Abstract — PoliSim è una piattaforma computazionale a tre stadi per la simulazione elettorale nel sistema misto proporzionale-uninominale italiano (Rosatellum bis). Lo Stage 1 applica uno swing model uniforme ai dati ufficiali del Ministero dell'Interno (Eligendo) su 221 collegi uninominali. Lo Stage 2 affina l'allocazione dei seggi attraverso un modello MRP (Multilevel Regression and Poststratification) implementato in PyMC 5.28.4 con verosimiglianza Dirichlet-Multinomial, calibrato su 252 osservazioni in 13 regioni. Lo Stage 3 genera strategie di comunicazione demograficamente mirate integrando profili psicografici derivati da ESS Round 11 (N=2.865), TRIPOL IT (N=1.231) e dati di calibrazione dalla Meta Ad Library. Il backtest su 14 collegi Camera del Lazio (elezioni politiche 2022) restituisce MAE 4,01 pp, RMSE 4,64 pp e winner accuracy 14/14. Il modello MRP supera una baseline OLS di 1,04 pp RMSE su sei set di validazione regionale. I limiti noti — tra cui copertura CI90 del 29%, bias CDX residuo di −3,33 pp e assenza di dati sondaggistici proprietari — sono esplicitamente dichiarati.

1. Introduzione e Motivazione

Il sistema elettorale italiano introdotto dalla Legge 165/2017 — noto come Rosatellum bis — combina allocazione proporzionale e maggioritaria: il 37% dei seggi in entrambi i rami del Parlamento è assegnato tramite collegi uninominali, mentre il restante 63% è distribuito proporzionalmente tra liste bloccate. Questa struttura ibrida genera interdipendenze non lineari tra quote di voto nazionali, forza dei candidati locali e composizione delle coalizioni, rendendo metodologicamente complessa la proiezione a livello di collegio.

Gli strumenti di previsione attualmente disponibili per analisti e operatori politici italiani si suddividono tipicamente in due categorie: (i) aggregatori di sondaggi a livello nazionale che riportano le intenzioni di voto senza tradurle in distribuzioni di seggi per collegio, oppure (ii) modelli proprietari dei principali network radiotelevisivi le

cui specifiche metodologiche non sono pubblicamente accessibili. Persiste dunque un divario tra i dati sondaggistici liberamente consultabili e proiezioni di seggio operativamente utili.

PoliSim colma questo divario attraverso una piattaforma riproducibile e a metodologia aperta articolata in tre stadi computazionali. La filosofia di progettazione si fonda su tre principi: trasparenza metodologica, con tutte le assunzioni di modellazione e i limiti noti esplicitamente dichiarati; fondamento empirico, utilizzando esclusivamente fonti istituzionali ufficiali (Eligendo, ISTAT, ESS) come input; e validazione progressiva, che richiede backtest multi-regionale prima di qualsiasi affermazione pubblica.

La piattaforma è stata sviluppata nell'ambito di Q-Italia (qitalia.org), un'organizzazione di civic technology che applica l'analisi politica quantitativa al dibattito pubblico italiano. Q-Italia utilizza PoliSim come caso di studio in tempo reale, dimostrando le sue capacità su scenari politici reali. Questa duplice funzione — strumento di ricerca e infrastruttura di comunicazione — informa i requisiti di progetto: la piattaforma deve essere al contempo accademicamente difendibile e operativamente dispiegabile.

1.1 Domande di Ricerca

PoliSim è progettata per rispondere a tre domande operative:

- Q1 — Allocazione dei seggi: Data una distribuzione del voto nazionale osservata o ipotetica, qual è la distribuzione attesa dei seggi nella Camera dei Deputati e nel Senato della Repubblica?

- Q2 — Verifica della doppia maggioranza: Uno scenario dato produce una maggioranza di governo stabile in entrambi i rami del Parlamento simultaneamente, tenendo conto della diversa composizione demografica dell'elettorato del Senato?

- Q3 — Ottimizzazione del messaggio: Dato uno scenario elettorale o di policy, quale strategia di comunicazione massimizza la portata persuasiva sui segmenti demografici definiti?

1.2 Perimetro e Limiti

Questo white paper documenta la metodologia, le fonti di dati e i risultati di validazione per PoliSim versione 2.1, in produzione a maggio 2026. La piattaforma opera come strumento di analisi di scenario, non come strumento sondaggistico predittivo. Non genera previsioni autonome delle intenzioni di voto; piuttosto, traduce distribuzioni di voto fornite esternamente (da aggregatori di sondaggi pubblici o scenari definiti dall'utente) in distribuzioni di seggi con stime di incertezza associate.

La distinzione è metodologicamente rilevante: PoliSim può modellare accuratamente le conseguenze elettorali di un dato scenario, ma non può prevedere autonomamente quale scenario si materializzerà. L'integrazione con aggregatori di sondaggi esterni (prevista come fase di sviluppo successiva) colmerebbe questo divario.

2. Fonti di Dati

PoliSim attinge a quattro fonti di dati primarie, ciascuna con un ruolo distinto nella pipeline di modellazione. Un principio cardine dell'architettura dati è il ricorso esclusivo a fonti istituzionali ufficiali con provenienza documentata, al fine di garantire la riproducibilità e resistere alle contestazioni di credibilità associate a dataset proprietari o non documentati.

2.1 Eligendo — Archivio Elettorale del Ministero dell'Interno

L'input primario dello Stage 1 è la banca dati ufficiale dei risultati elettorali gestita dal Ministero dell'Interno attraverso la piattaforma OpenData Eligendo. Per le elezioni politiche del 2022, i dati coprono tutti i 221 collegi uninominali a livello di collegio (Camera: 147 collegi; Senato: 74 collegi), inclusi voti per partito, composizione delle coalizioni, nomi dei candidati e conteggi ufficiali.

I dati vengono scaricati come archivi ZIP dal portale Eligendo e processati da `polisim*build*collegi.py`, che analizza i risultati a livello di collegio e costruisce la baseline dello swing model. Un problema persistente di qualità dei dati — documentato e gestito, non occultato — è l'inconsistenza della codifica dei nomi dei comuni tra versioni del dataset (mojibake UTF-8/Latin-1), in particolare per i comuni bilingui in Alto Adige e Valle d'Aosta. È stato ottenuto un tasso di corrispondenza del 98,4% sui nomi dei comuni; i record non abbinati sono esclusi dal join geografico ma conservati nei totali aggregati.

Anomalie nelle coalizioni locali — in particolare la Südtiroler Volkspartei (SVP) in Alto Adige e il gruppo Misto in diversi collegi — introducono circa 17 seggi di discrepanza tra i totali calcolati e i risultati CDX documentati su Wikipedia. Questa discrepanza è documentata, spiegata, e non inficia la validità dello swing model per il 97% dei collegi in cui si applicano le composizioni di coalizione standard.

2.2 Censimento Decennale ISTAT 2021

Le variabili demografiche per il modello MRP (Stage 2) derivano dal Censimento decennale ISTAT 2021. I dati sono ottenuti tramite download massivo dei file indicatori a livello di sezione (R12*indicatori*2021_sezioni.xlsx per il Lazio, contenente 39.670 sezioni in cinque province) e aggregati al livello di collegio tramite join geografico con i confini ufficiali degli shapefile (Collegi Elettorali Basi Geografiche, MIT).

Otto variabili demografiche sono attualmente in uso nel modello MRP: percentuale di residenti di età 15-34 anni (pct*giovani), percentuale di età 65+ (pct*anziani), percentuale con istruzione terziaria (pct*alta*istr), percentuale femminile (pct*femmine), tasso di occupazione (pct*occupati), percentuale di nati all'estero (pct*stranieri) e percentuale di origine extra-UE (pct*extra*ue), più un indicatore del tipo di elezione (*is*politiche).

Un limite strutturale dell'implementazione attuale è che le variabili ISTAT sono completamente popolate solo per i 14 collegi laziali; i restanti 238 collegi ricevono un valore zero per le variabili legate all'immigrazione (pct*stranieriz*, pct*extra*ue_z). Sebbene questo sia un limite riconosciuto, non introduce distorsione per i collegi extra-Lazio, poiché l'imputazione a zero è uniforme e il modello apprende intercette specifiche per regione attraverso la struttura gerarchica. L'estensione della copertura ISTAT a tutti i 252 collegi costituisce una milestone di sviluppo prioritaria.

Variabile	Descrizione	Fonte	Copertura
pct_giovani	% residenti 15–34 anni	ISTAT 2021	Tutti 252 collegi
pct_anziani	% residenti 65+	ISTAT 2021	Tutti 252 collegi
pct_alta_istr	% con istruzione terziaria	ISTAT 2021	Tutti 252 collegi
pct_femmine	% residenti femminili	ISTAT 2021	Tutti 252 collegi
pct_occupati	Tasso di occupazione	ISTAT 2021	Tutti 252 collegi
pct_stranieri	% nati all'estero	ISTAT 2021	Solo 14 Lazio*
pct_extra_ue	% di origine extra-UE	ISTAT 2021	Solo 14 Lazio*
is_politiche	Binario: politiche (1) vs regionali (0)	Eligendo	Tutti 252 collegi

** I collegi non laziali ricevono valore imputato pari a zero. L'estensione alla copertura nazionale è in programma.*

2.3 Dati d'Indagine — ITANES, ESS Round 11, TRIPOL IT

Lo Stage 3 (message optimizer) attinge a tre dataset di indagine complementari per costruire profili psicografici per dieci segmenti demografici:

- ESS Round 11 (2023–24): N=2.865 intervistati italiani, campione probabilistico face-to-face. Variabili utilizzate: fiducia istituzionale, auto-collocazione sinistra-destra, atteggiamenti sull'immigrazione, soddisfazione democratica. I dati rappresentano valori strutturali post-elezioni 2022; non vengono usati come proxy delle intenzioni di voto.

- TRIPOL IT (2021–22): N=1.231 intervistati italiani. Variabili: punteggi di polarizzazione affettiva, indice dei valori politici WAPSV. Utilizzato per calibrare la valenza emotiva della comunicazione politica tra i segmenti.

- Meta Ad Library (maggio 2026): 30 inserzioni politiche e di ONG analizzate (STC, UNICEF, Emergency). Utilizzata esclusivamente per la calibrazione del tono e dell'hook nei segmenti donatori lasciti e responsabili CSR. Questa fonte non viene usata per l'inferenza demografica.

Limite dichiarato: la copertura TRIPOL del segmento TERZO_POLO si basa su N=44 intervistati ESS — un sottocampione statisticamente fragile. Le stime per questo segmento devono essere interpretate con incertezza elevata. Il dataset ITANES (Indagine Nazionale sugli Elettori Italiani) fornisce ulteriore contesto di validazione longitudinale ma non viene utilizzato come input diretto del modello.

2.4 Risultati Elettorali Regionali — Dataset di Validazione

La validazione del modello attinge ai risultati ufficiali Eligendo di 13 regioni italiane in più cicli elettorali (2020–2023), per un totale di 252 osservazioni collegio-elezione. Sei regioni sono state selezionate come benchmark di validazione primario per il confronto OLS vs. MRP: Toscana, Emilia-Romagna, Veneto, Sicilia, Sardegna e Puglia. I 14 collegi del Lazio (Camera uninominale, politiche 2022) costituiscono il dataset di backtest primario, scelto perché i dati ISTAT sono disponibili a piena risoluzione sezione per questa regione.

3. Stage 1 — Swing Model Elettorale

Lo Stage 1 implementa uno swing model uniforme a livello di collegio applicato ai 221 seggi uninominali del sistema Rosatellum. Il modello traduce una distribuzione del voto nazionale specificata dall'utente o proveniente da fonte esterna in esiti a livello di collegio propagando le variazioni della quota di voto per coalizione rispetto alla baseline 2022.

3.1 Specificazione del Modello

Sia V^{2022}*c la quota di voto osservata della coalizione c nelle elezioni del 2022 e V^sc*c la quota di voto dello scenario specificato dall'utente. Lo swing a livello di coalizione è definito come:

$\Delta_c = V^{sc}_c - V^{2022}_c$

Per ciascun collegio uninominale i, il vincitore in carica c*_i mantiene il seggio a meno che il vantaggio cumulato dello sfidante più forte non superi la metà del margine di vittoria osservato nel 2022:

seat_i ← c_sfidante se $\max_{c \neq c^*}(\Delta_c - \Delta_{c^*}) > \text{margine}_i / 2$

seat_i ← c* altrimenti

Il margine di vittoria margine_i è desunto direttamente dai risultati ufficiali Eligendo 2022 per i 14 collegi Camera del Lazio, e dalle medie di macro-area per i restanti 207 collegi. Questa asimmetria — margini reali per il Lazio, margini aggregati altrove — riflette lo stato attuale della disponibilità delle variabili ISTAT e costituisce la principale approssimazione nota dello Stage 1.

3.2 Allocazione Proporzionale

Il 63% dei seggi allocato proporzionalmente (245 Camera; 122 Senato) è distribuito con il metodo d'Hondt con soglia nazionale al 3% (Rosatellum Art. 83, c. 1(c)). L'implementazione applica la soglia partito per partito, poi ridimensiona i totali dei partiti ammessi al 100% prima dell'allocazione:

ammessi = {p : V_p ≥ 3,0% e p qualifica come lista}

seggi_p = floor(V_p / sum(V_ammessi) × N_seggi) + correzione resti

I seggi residui sono assegnati per il resto frazionario più elevato (metodo Hamilton/Hare). I totali di coalizione nel segmento proporzionale sono calcolati sommando le allocazioni per partito secondo la mappa di appartenenza alle coalizioni dichiarata (MAPPA_COAL), che attualmente copre CDX, CSX, M5S e CENTRO. I partiti non assegnati a nessuna coalizione confluiscono nella categoria ALTRI.

3.3 Struttura delle Macro-Aree

Per i 207 collegi non laziali, lo Stage 1 opera su un'aggregazione per macro-area derivata da Eligendo 2022. Le cinque macro-aree (Nord-Ovest, Nord-Est, Centro, Sud, Sud-Isole) sono parametrizzate da: numero di collegi, vincitori 2022 per coalizione, margine di vittoria medio e numero di collegi contendibili ("collegi in bilico").

Macro-area	Collegi Camera	CDX 2022	CSX 2022	M5S 2022	Margine medio	In bilico
Nord-Ovest	38	28	5	0	12,5 pp	8
Nord-Est	27	22	3	0	14,2 pp	5
Centro*	14	14	0	0	24,9 pp	4
Sud	36	22	4	8	8,1 pp	14
Sud-Isole	18	8	2	7	6,4 pp	9

** La macro-area Centro è parzialmente sostituita dai dati a livello di collegio del Lazio nella v2.1; i restanti collegi del Centro Italia utilizzano i parametri di macro-area.*

Il numero di seggi che cambiano mano in una macro-area è limitato superiormente dal numero di collegi in bilico ed è funzione lineare del vantaggio di swing:

seggi_cambiati = min(n_bilico × Δ_vantaggio / (margine_medio + ε), n_vinti_2022)

Questa forma funzionale sottostima deliberatamente i cambiamenti di seggio in scenari di tracollo — una scelta conservativa che riduce la falsa precisione. L'uso appropriato dello Stage 1 è l'analisi di scenario vicina alla baseline (swing ±5–8 pp); gli scenari estremi devono essere interpretati con cautela aggiuntiva.

3.4 Verifica della Doppia Maggioranza

Lo Stage 1 calcola simultaneamente i totali di seggi per entrambe le camere e verifica se una coalizione o combinazione di coalizioni ottiene la maggioranza di governo (Camera: 201 seggi; Senato: 104 seggi). L'output segnala esplicitamente tre condizioni: doppia*maggioranza (maggioranza in entrambi i rami), maggioranza*parziale (maggioranza in un solo ramo) e nessuna maggioranza. Questa funzionalità risponde direttamente al vincolo costituzionale distintivo del sistema italiano, in cui un governo deve mantenere la fiducia dei due rami del Parlamento in modo indipendente.

4. Stage 2 — Multilevel Regression and Poststratification

Lo Stage 2 applica un modello MRP (Multilevel Regression and Poststratification) per affinare la distribuzione dei seggi prodotta dallo Stage 1, tenendo esplicitamente conto dell'eterogeneità demografica tra collegi. L'MRP è stato originariamente sviluppato da Gelman e Little (1997) e sistematicamente valutato per la previsione elettorale da Wang et al. (2015), che hanno dimostrato un MAE di circa 2–3 pp nelle stime delle quote di voto presidenziali a livello statale negli USA. PoliSim adatta il framework al contesto italiano multipartitico e bicamerale.

4.1 Architettura del Modello

Il modello MRP è implementato in PyMC 5.28.4 (Salvatier et al., 2016). Gli esiti osservati sono i risultati elettorali dei registri ufficiali Eligendo su 252 osservazioni collegio-elezione (13 regioni, più cicli elettorali 2020–2023). Il modello stima simultaneamente la quota di voto per quattro coalizioni.

4.2 Funzione di Verosimiglianza

Per modellare la distribuzione congiunta delle quote di voto tra quattro coalizioni concorrenti si utilizza una verosimiglianza Dirichlet-Multinomial. Questa scelta è deliberata e risponde a due requisiti specifici: (i) le quote di voto devono sommare a uno tra le coalizioni, e (ii) il modello deve trattare tutte le coalizioni simmetricamente senza privilegiare nessun competitore. Specificazioni precedenti basate su una verosimiglianza Normale per i margini CDX introducevano distorsioni asimmetriche; il passaggio al Dirichlet-Multinomial è stato motivato esplicitamente dall'overprediction in direzione CDX osservata nelle sessioni di validazione preliminari.

Per il collegio i con totale voti N*i, i conteggi di voto osservati (y*{i,CDX}, y*{i,CSX}, y*{i,M5S}, y_{i,CENTRO}) sono modellati come:

y_i | α_i, N_i ~ DirichletMultinomial(N_i, α_i)

α_{i,c} = softmax(μ_c + Σ_k β_{k,c} × X_{ik} + γ_{regione[i],c})

dove μ*c è l'intercetta specifica per coalizione, β*{k,c} è il coefficiente della k-esima variabile demografica per la coalizione c, X*{ik} è il valore standardizzato della variabile ISTAT, e γ*{regione,c} è un effetto casuale specifico per regione.

4.3 Set di Variabili e Distribuzioni a Priori

Le otto variabili utilizzate nel trace v4 sono elencate di seguito con le rispettive distribuzioni a priori. Tutte le variabili continue sono standardizzate (z-score) sul set di addestramento di 252 osservazioni prima dell'adattamento del modello:

Variabile (z-score)	Descrizione	Distribuzione a priori
pct_giovani_z	% residenti 15–34 anni	Normal(0, 1)
pct_anziani_z	% residenti 65+	Normal(0, 1)
pct_alta_istr_z	% con istruzione terziaria	Normal(0, 1)
pct_femmine_z	% residenti femminili	Normal(0, 1)
pct_occupati_z	Tasso di occupazione	Normal(0, 1)
pct_stranieri_z	% nati all'estero (solo Lazio)	Normal(0, 1)
pct_extra_ue_z	% di origine extra-UE (solo Lazio)	Normal(0, 1)
is_politiche	Binario: politiche (1) vs regionali (0)	Bernoulli implicito

Tutte le distribuzioni a priori Normal(0,1) riflettono prior debolmente informativi centrati sull'effetto nullo. Il parametro di concentrazione Dirichlet utilizza un iperprior Half-Normal(1).

La razionale della selezione delle variabili pct*stranieriz* e pct*extra*ue_z (aggiunte nel trace v4) è stata stabilita attraverso un'analisi sistematica di backtest: su 14 collegi del Lazio, queste due variabili hanno mostrato la correlazione più forte con l'errore di predizione MRP (r = −0,800 e r = −0,792 rispettivamente), superando tutte le altre variabili candidate come predittori di bias. La loro inclusione nella v4 ha ridotto il MAE di 0,66 pp rispetto alla v3 (da 4,67 a 4,01 pp).

4.4 Struttura Gerarchica ed Effetti Regionali

Gli effetti casuali a livello regionale γ_{regione,c} catturano le differenze politiche strutturali tra le macro-regioni italiane (ad esempio, il tradizionalmente forte voto M5S nelle regioni meridionali, il predominio CDX nel Nord-Est). Questi effetti sono modellati con un iperprior half-Cauchy debolmente informativo sulla deviazione standard:

γ_{regione,c} ~ Normal(0, σ_regione)

σ_regione ~ HalfCauchy(1)

La variabile binaria is_politiche tiene conto della differenza sistematica tra le elezioni politiche nazionali (dove il Rosatellum si applica integralmente) e le elezioni regionali (strutture dei candidati diverse, affluenza più bassa, quote M5S e liste locali più elevate in alcune regioni). L'inclusione delle elezioni regionali nel set di addestramento amplia la N da 14 (solo Lazio 2022) a 252, migliorando sostanzialmente le basi statistiche delle stime.

4.5 Configurazione del Campionamento

Il trace v4 è stato stimato con NUTS (No U-Turn Sampler) con la seguente configurazione:

draws=2000, tune=1000, chains=4, target_accept=0,90 (default)

Diagnostiche di convergenza: r*hat massimo = 1,000 su tutti i parametri (criterio di Gelman-Rubin; soglia: r*hat < 1,01); 125 transizioni divergenti (0,16% dei campioni post-warmup). Il numero di divergenze è aumentato da 10 (v3) a 125 (v4) in seguito all'aggiunta di due nuove variabili, indicando una lieve complessità della geometria a posteriori. A 125/8000 campioni post-warmup (1,6%), le divergenze rientrano nell'intervallo generalmente considerato accettabile per il lavoro applicato, ma richiedono monitoraggio nelle future iterazioni del modello. È prevista una sessione di tuning con target_accept=0,99 come miglioramento post-release.

5. Validazione e Backtesting

La validazione del modello MRP di PoliSim segue un protocollo a due livelli: (i) backtest a livello di collegio su 14 collegi Camera del Lazio utilizzando i risultati delle politiche 2022 come riferimento held-out, e (ii) confronto OLS vs. MRP a livello regionale su sei regioni italiane. Entrambi i protocolli utilizzano esclusivamente previsioni out-of-sample — il modello è addestrato sull'intero dataset di 252 osservazioni e le previsioni sono generate tramite cross-validazione leave-one-region-out per il confronto OLS e tramite previsione diretta per il backtest laziale.

5.1 Backtest Lazio — 14 Collegi

Il benchmark di validazione primario è il backtest su 14 collegi Camera uninominali del Lazio (politiche 2022). Questa regione è stata selezionata come dataset di validazione perché i dati del Censimento ISTAT 2021 sono disponibili a piena risoluzione sezione, consentendo una genuina previsione demografica MRP piuttosto che stime con imputazione a zero. I risultati di seguito riportano le previsioni del modello MRP v4 a fronte dei conteggi ufficiali Eligendo 2022.

Metrica	MRP v3	MRP v4	Interpretazione
MAE quota voto CDX	4,67 pp	4,01 pp	↓ miglioramento del 14,1%
RMSE quota voto CDX	6,08 pp	4,64 pp	↓ miglioramento del 23,7%
Bias (errore medio) CDX	−3,69 pp	−3,33 pp	Sottostr. sistemica, ridotta
Winner accuracy	14/14	14/14	100% collegi vinti correttamente
Copertura empirica CI90	29%	29%	⚠ Problema strutturale (cfr. §6)
Max P(CDX vince) in seggi CDX	n/d	100%	Previsioni ad alta confidenza

CDX = coalizione Centrodestra (FdI, Lega, FI). La quota di voto è percentuale del totale voti validi. La copertura CI90 indica la frazione di esiti reali che ricadono nell'intervallo credibile a posteriori al 90%.

I risultati a livello di collegio rivelano un pattern geografico sistematico nei residui: i cinque collegi di Lazio 2 (prevalentemente provinciali/rurali) mostrano errori assoluti più elevati (errore medio |e| = 3,1 pp) rispetto ai nove collegi di Lazio 1 centrati su Roma (|e| = 5,2 pp per la periferia romana, |e| = 0,97 pp per Roma centrale). Questa eterogeneità suggerisce che il set attuale di variabili cattura meglio i gradienti demografici urbani rispetto alle dinamiche provinciali, e motiva l'inclusione di variabili aggiuntive (composizione del mercato del lavoro da DCSC_CONDPROFOCCUP) nella specificazione v5 pianificata.

5.2 Confronto con Baseline OLS — Sei Regioni

Per valutare se il modello MRP fornisce un genuino miglioramento rispetto a una baseline statistica più semplice, è stata stimata una regressione OLS utilizzando la stessa matrice di progetto a 8 variabili sugli stessi dati di addestramento. L'RMSE out-of-sample è stato calcolato per sei regioni held-out su singoli cicli elettorali:

Regione	Elezione	RMSE OLS	RMSE MRP	**Δ** (vantaggio MRP)
Toscana	Reg. 2020	8,21 pp	7,43 pp	+0,78 pp
Emilia-Romagna	Reg. 2020	9,84 pp	8,67 pp	+1,17 pp
Veneto	Reg. 2020	7,92 pp	7,31 pp	+0,61 pp
Sicilia	Reg. 2022	10,12 pp	8,91 pp	+1,21 pp
Sardegna	Reg. 2024	8,76 pp	8,10 pp	+0,66 pp
Puglia	Reg. 2020	8,94 pp	6,54 pp	+2,40 pp
Media (6 regioni)	—	8,96 pp	7,82 pp	+1,14 pp

La baseline OLS utilizza il set identico di variabili e dati di addestramento. Vantaggio MRP = RMSE_OLS − RMSE_MRP. Valori positivi indicano che MRP supera OLS. Tutte le elezioni held-out non erano incluse nel set di addestramento per quel fold.

Il modello MRP supera la baseline OLS in tutte e sei le regioni held-out, con un vantaggio medio di RMSE di 1,14 pp. Il vantaggio è massimo in Puglia (+2,40 pp), regione con elevata eterogeneità demografica tra le aree urbane costiere e le circoscrizioni agricole dell'entroterra, coerente con l'aspettativa che la modellazione gerarchica fornisca il maggior beneficio laddove l'eterogeneità intra-regionale è elevata.

Per riferimento, Wang et al. (2015) riportano MAE di circa 2–3 pp per le stime della quota di voto presidenziale USA da modelli MRP addestrati su sondaggi online opt-in. L'RMSE di PoliSim di 7,82 pp sulle elezioni regionali out-of-sample è sostanzialmente più elevato, riflettendo due ulteriori fonti di difficoltà: (i) il contesto italiano multipartitico, in cui gli errori di previsione si cumulano su quattro coalizioni anziché su un esito binario, e (ii) l'assenza di dati sondaggistici proprietari a livello di collegio, che Wang et al. utilizzano come input chiave. Il benchmark PoliSim da una PoC di regressione held-out sul Lazio (RMSE 3,91 pp) dimostra che il modello può avvicinarsi all'accuratezza di Wang et al. quando sono disponibili dati demografici ad alta risoluzione.

6. Limitazioni Dichiarate

La credibilità scientifica richiede che i limiti siano dichiarati con la stessa prominenza dei risultati. Le seguenti limitazioni sono proprietà architetturali dell'implementazione attuale di PoliSim — non lacune provvisorie da correggere silenziosamente, ma caratteristiche strutturali dell'approccio di modellazione che influenzano l'interpretazione di tutti gli output. Ci si aspetta che gli utenti della piattaforma abbiano letto questa sezione.

6.1 Mancata Copertura degli Intervalli di Confidenza

Il limite noto più significativo del modello MRP è la copertura empirica CI90 del 29% sui 14 collegi del Lazio, contro l'attesa nominale del 90%. Ciò significa che in circa il 71% dei casi il vero esito cade al di fuori dell'intervallo credibile a posteriori al 90% — un grave fallimento della calibrazione degli intervalli.

Questo non indica che le stime puntuali siano errate (MAE 4,01 pp è operativamente utile), ma che gli intervalli di incertezza riportati dal modello sono drammaticamente sovra-confidenti e non devono essere usati per l'inferenza probabilistica. La causa principale è la varianza a posteriori insufficiente: la verosimiglianza Dirichlet-Multinomial con la struttura a priori attuale concentra la massa di probabilità più strettamente di quanto la distribuzione campionaria reale richieda. La soluzione richiede prior più ricchi sul parametro di concentrazione o una componente esplicita di inflazione della varianza — non variabili aggiuntive.

Indicazione operativa: Gli output CI dello Stage 2 devono essere interpretati come indicatori direzionali di incertezza, non come affermazioni probabilistiche calibrate. Le stime puntuali (MAP) sono gli output

operativamente validi. Le affermazioni probabilistiche ("X ha una probabilità Y% di vincere") non devono essere derivate dall'attuale struttura CI senza ricalibrazione della varianza.

6.2 Bias Sistematico CDX

Un errore medio di previsione di −3,33 pp per la quota di voto CDX persiste nel trace v4 (v3: −3,69 pp). Il modello sottostima sistematicamente le performance del CDX. L'analisi dei residui per tipo di collegio rivela che il bias è concentrato nei collegi di Lazio 2 (provinciali, prevalentemente piccoli comuni), dove il CDX ha superato le previsioni del modello in media di 4,8 pp. I collegi urbani di Lazio 1 mostrano un bias sistematico inferiore (errore medio −2,1 pp).

Il bias è parzialmente spiegabile da variabili non ancora presenti nel modello — in particolare la quota di occupazione agricola e i pattern di voto strutturali dei piccoli comuni, che non sono catturati dalle attuali variabili del Censimento ISTAT 2021. Il dataset DCSC_CONDPROFOCCUP (occupazione per settore ISTAT a livello comunale) è identificato come la fonte di dati di massima priorità per la riduzione del bias nella specificazione v5 pianificata. Il bias non incide sull'accuratezza winner agli attuali livelli di predominio CDX nel Lazio, ma diventerebbe consequenziale in scenari di quasi-parità.

6.3 Assenza di Dati Sondaggistici Proprietari

PoliSim è un *modello ecologico*: stima le quote di voto da variabili demografiche e storiche, non da risposte a sondaggi individuali legate a localizzazioni geografiche. Questo è il limite strutturale definitorio dell'approccio. Il modello non può "sapere" che il consenso di un partito si è spostato in una particolare regione a meno che tale spostamento non sia rilevabile dal registro storico Eligendo o dal profilo demografico ISTAT.

L'implicazione pratica è che PoliSim non può prevedere autonomamente quale scenario elettorale si materializzerà. È un *calcolatore di scenari ipotetici*, non uno strumento predittivo. L'integrazione pianificata con dati aggregati di sondaggi esterni (supermedie Politpro, attualmente ingerite tramite scraper settimanale via cron) consente alla piattaforma di rispondere a "dati i sondaggi di oggi, quale sarebbe la distribuzione dei seggi?" — ma questo rimane contingente all'accuratezza dei dati sondaggistici esterni.

6.4 Copertura ISTAT delle Variabili

Come documentato nella Sezione 2.2, le variabili pct*stranieriz* e pct*extra*ue_z sono popolate con dati reali solo per 14 dei 252 collegi. I restanti 238 collegi ricevono un'imputazione a zero per queste variabili. Ciò significa che il modello completo a 8 variabili è effettivamente un modello a 6 variabili per il 94% dei collegi, riducendo i guadagni di accuratezza a quelli attribuibili alle sei variabili a copertura universale.

L'estensione della copertura ISTAT a tutti i 252 collegi richiede il download e il processamento dei file del Censimento a livello di sezione per tutte e 20 le regioni italiane (contro il solo Lazio). I dati esistono e sono pubblicamente accessibili; il lavoro è un'attività ingegneristica stimata in 5–8 sessioni di sviluppo. Fino al suo completamento, i miglioramenti di RMSE derivanti dalle variabili v4 devono essere intesi come risultati specifici del Lazio che potrebbero non generalizzarsi a livello nazionale.

6.5 Stage 3 — Pesi di Calibrazione Non Validati

Il message optimizer (Stage 3) genera strategie di comunicazione demograficamente mirate applicando pesi di calibrazione expert-calibrated ai dati di indagine ESS Round 11 e TRIPOL IT. Questi pesi determinano quali frame di messaggio sono prioritari per ciascuno dei dieci segmenti demografici. I pesi sono stati calibrati da esperti di dominio sulla base della revisione dei dati di performance della Meta Ad Library (30 inserzioni) e della revisione qualitativa dei microdati di indagine, ma non sono ancora stati validati tramite A/B test controllati su esiti reali di campagna.

La conseguenza operativa è che gli output dello Stage 3 devono essere trattati come ipotesi strutturate da sottoporre a test, non come raccomandazioni empiricamente validate. È stato progettato un protocollo di validazione che richiede 20 campagne partner con misurazione standardizzata degli esiti, attualmente in fase di attivo reclutamento.

7. Stage 3 — Message Optimizer

Lo Stage 3 genera strategie di messaggio per campagne politiche, organizzazioni civiche e ONG, calibrate sul profilo demografico e psicografico di ciascun segmento target. Il componente operazionalizza la teoria della comunicazione politica — in particolare la teoria del framing di Entman (1993) e i successivi sviluppi empirici nella agenda-setting e

nella comunicazione politica mirata (Nickerson & Rogers, 2010; Broockman & Kalla, 2016) — attraverso una pipeline strutturata di prompt engineering costruita sull'API Claude di Anthropic.

7.1 Architettura dei Profili Psicografici

Per il contesto politico italiano sono definiti dieci segmenti demografico-psicografici. Ciascun segmento è caratterizzato da un profilo multidimensionale che combina descrittori sociodemografici oggettivi (età, istruzione, condizione lavorativa, area geografica) con variabili attitudinali e comportamentali derivate da fonti di indagine:

Segmento	N (base survey)	Fonte primaria	Assi attitudinali chiave
Giovani precari Sud	N≈380 ESS R11	ESS R11 + TRIPOL	Sfiducia istituzionale, intento emigratorio
Casalinghe disoccupate Sud	N≈290 ESS R11	ESS R11	Riconoscimento lavoro di cura, accesso sanità
Operai/artigiani Nord	N≈310 ESS R11	ESS R11 + TRIPOL	Rischio automazione, pressione fiscale, diffidenza sinistra
Laureati urbani progressisti	N≈420 ESS R11	ESS R11	Qualità democratica, clima, costo affitti
Astensionisti valoriali	N≈180 ESS R11	ESS R11 + TRIPOL	Cinismo sistemico, discontinuità radicale
Pensionati Centro-Nord	N≈340 ESS R11	ESS R11	Sistema sanitario, sicurezza, fedeltà partitica
Donatori lasciti testamentari	N≈85 ESS R11	ESS R11 + Meta Ads	Eredità, fiducia istituzionale, tabù della morte
Responsabili CSR	Expert-calibrated	Meta Ads library	Accountability ESG, ROI reputazionale
Terzo Polo moderati	N≈44 ESS R11	ESS R11	⚠ Sottocampione esiguo — incertezza elevata
Imprenditori PMI	N≈95 ESS R11	ESS R11 + TRIPOL	Onere regolatorio, accesso al credito

ESS R11 = European Social Survey Round 11 (2023-24), sottocampione Italia N=2.865. TRIPOL = studio italiano sulla polarizzazione affettiva (2021-22), N=1.231. I valori N sono dimensioni approssimative del sottocampione per segmento.

7.2 Integrazione dei Dati Empirici

La costruzione dei profili segue un protocollo di integrazione dati gerarchico. L'ESS Round 11 fornisce la baseline attitudinale strutturale: fiducia istituzionale (scala 0–10), auto-collocazione sinistra-destra (scala 0–10), indice di atteggiamento verso l'immigrazione (composito) e soddisfazione democratica. Questi sono calcolati come medie e deviazioni standard a livello di segmento dal sottocampione italiano ESS, aggregati per celle età × istruzione × condizione lavorativa allineate con le definizioni di segmento.

TRIPOL IT contribuisce il layer della polarizzazione affettiva: punteggi WAPSV (World Values Survey Political Values) per segmento, e differenziali di favorabilità ingroup/outgroup. ESS Round 11 è post-elezioni 2022; TRIPOL copre il ciclo 2021–22. Entrambi i dataset sono utilizzati come proxy dei valori politici strutturali piuttosto che dell'intenzione di voto attuale — un vincolo interpretivo critico documentato in ogni output dello Stage 3.

Calibrazione Meta Ad Library: 30 inserzioni italiane politiche e di ONG sono state analizzate dalla Meta Ad Library (maggio 2026), coprendo le organizzazioni STC (aiuto umanitario), UNICEF Italia ed Emergency. Per i segmenti lasciti testamentari e responsabili CSR, i pattern di performance delle inserzioni (segnali di engagement, dati di sovrapposizione del pubblico) sono stati utilizzati per calibrare i parametri di tono e hook. Questa fonte è utilizzata esclusivamente per la calibrazione dello stile comunicativo; non viene usata per la profilazione demografica.

7.3 Architettura del Prompt e Struttura dell'Output

La pipeline dello Stage 3 inietta i dati empirici di profilo — dati*empirici — direttamente nel prompt di generazione per tutti i dieci segmenti in fase di esecuzione, caricati dal JSON del profilo ITANES enriched (itanes*profili*enriched.json). Per i due segmenti adiacenti alle ONG (donatori*lasciti, responsabili*csr), viene iniettato un ulteriore blocco calibrazione*meta_ads.

La struttura dell'output per ciascuna richiesta di ottimizzazione del messaggio comprende: (i) un frame di messaggio primario con citazione esplicita del trigger demografico, (ii) un frame secondario di "risposta all'obiezione" che affronta le barriere comunicative documentate del segmento, (iii) una raccomandazione di canale (piattaforma, formato, registro tonale), e (iv) una lista esplicita di elementi "da evitare" derivata dal profilo dei trigger negativi del segmento. Tutti gli output includono una nota che i pesi sono expert-calibrated e non sono ancora stati validati su esiti di A/B test.

7.4 Adattamento Commerciale

L'architettura dei segmenti politici è in corso di estensione per tre casi d'uso non politici: comunicazione di brand (aggiunta degli assi di scoring propensione*acquisto, premium*accessibile, razionale*emotivo), lasciti testamentari e donazioni maggiori per ONG (fiducia*istituzionale, eredità, delicatezza*morte), e raccolta fondi corporate per ONG (ritorno*reputazionale, ESG*alignment, misurabilità*impatto). Queste estensioni richiedono modifiche allo schema dell'output JSON ma non richiedono modifiche ai componenti MRP o swing model. La logica di segmentazione demografica si trasferisce direttamente dai contesti politici a quelli civici.

Riferimenti Bibliografici

Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, 352(6282), 220–224.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4), 51–58.

Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23(2), 127–135.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472.

Ministero dell'Interno — Repubblica Italiana (2022). Eligendo — Archivio storico delle elezioni. Disponibile su: https://elezioni.interno.gov.it/opendata

ISTAT (2021). 15° Censimento generale della popolazione e delle abitazioni — Indicatori per sezione di censimento. Dataset R12 Lazio.

Nickerson, D. W., & Rogers, T. (2010). Do you have a voting plan? Implementation intentions, voter turnout, and organic plan making. *Psychological Science*, 21(2), 194–199.

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, 12(4), 375–385.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2, e55.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3), 980–991.

PoliSim White Paper Tecnico v1.0 — Maggio 2026

polisim.dev | *admin@qitalia.org* | *github.com/AlCap27/polisim*