

NOTE TECHNIQUE

Méthodologie mathématique

Modèle probabiliste appliqué aux sondages présidentiels 2027

Ce document expose la procédure utilisée pour transformer des sondages publics en probabilités de qualification et de victoire. Il présente d'abord les principes de lecture, puis les définitions mathématiques nécessaires à l'audit du modèle : pondération des enquêtes, incertitude, simulations de premier tour, duels de second tour, pondération des scénarios et agrégation finale.

Objet	Agrégation probabiliste de sondages électoraux.
Périmètre	Premier tour par scénario, duels de second tour, scénarios de candidatures, probabilité totale de victoire.
Nature	Document méthodologique de référence.
Version	1.0.
Date	1 juin 2026.
Nombre de simulations	100 000 tirages Monte Carlo par étape simulée.

Table des matières

1	Objet et convention de lecture	1
2	Architecture générale	1
3	Pondération des sondages	1
3.1	Poids temporel et poids d'échantillon	1
3.2	Score central	2
4	Incertitude	2
4.1	Erreur d'échantillonnage	2
4.2	Dispersion empirique	2
4.3	Plancher d'incertitude	2
4.4	Écart-type nowcast	2
5	Projection temporelle	3
6	Premier tour	3
6.1	Traitement par scénario	3
6.2	Probabilités de rang	3
7	Second tour	4
8	Pondération des scénarios	4
8.1	Événements structurés	4
8.2	Score de plausibilité	4
8.3	Softmax	4
9	Probabilité totale de victoire	5
10	Précision numérique	5
11	Limites	5

1 Objet et convention de lecture

Le modèle ne cherche pas à produire une certitude électorale. Il construit une distribution de trajectoires compatibles avec les sondages disponibles, les hypothèses de candidatures et les duels de second tour effectivement observés dans les données.

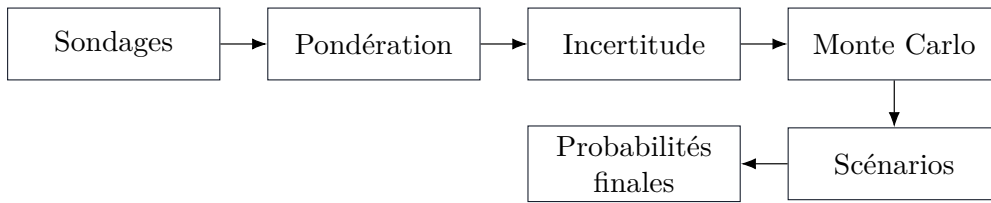
Convention de lecture

Une probabilité publiée doit être comprise comme une fréquence de réalisation dans les simulations du modèle. Elle est conditionnelle aux données intégrées, aux scénarios retenus et aux paramètres d'incertitude décrits ci-dessous.

Cette convention est essentielle. Une valeur de 80 % ne signifie pas que l'événement est acquis ; elle signifie que, dans l'univers formalisé par les hypothèses du modèle, l'événement se produit dans environ quatre simulations sur cinq.

2 Architecture générale

La chaîne de calcul est volontairement séquentielle. Chaque étape produit un objet mathématique utilisé par l'étape suivante.



On note i un sondage, c un candidat, s un scénario de premier tour, d un duel de second tour. Les scores sont exprimés en points de pourcentage.

3 Pondération des sondages

3.1 Poids temporel et poids d'échantillon

Soit T la date d'analyse et t_i la date de fin de terrain du sondage i . Son âge est

$$a_i = \max(0, T - t_i).$$

La fraîcheur est modélisée par une décroissance exponentielle de demi-vie 90 jours :

$$r_i = 0.5^{a_i/90}.$$

Si n_i est la taille d'échantillon, le poids utilisé dans les moyennes est

$$w_i = \sqrt{n_i} r_i.$$

Interprétation

Un sondage récent et large reçoit davantage de poids. La racine carrée évite toutefois qu'une enquête de très grande taille domine mécaniquement l'ensemble : doubler l'échantillon améliore le poids d'un facteur $\sqrt{2}$, et non d'un facteur 2. Ce choix reflète l'idée statistique classique selon laquelle la précision d'une proportion progresse comme la racine carrée de la taille d'échantillon. Un sondage ancien n'est pas supprimé brutalement : son influence décroît progressivement.

3.2 Score central

Pour un ensemble I de sondages comparables, le score central est la moyenne pondérée

$$\bar{x} = \frac{\sum_{i \in I} w_i x_i}{\sum_{i \in I} w_i}.$$

Ce score est une estimation de l'état courant de l'opinion dans l'hypothèse considérée. Il ne constitue pas encore une projection électorale complète.

4 Incertitude

4.1 Erreur d'échantillonnage

On pose $p_i = x_i/100$. L'écart-type binomial théorique, corrigé par un facteur de plan d'enquête égal à 1.5, est

$$\sigma_{\text{ech},i} = 100 \sqrt{\frac{p_i(1-p_i)}{n_i}} \sqrt{1.5}.$$

Après agrégation, l'erreur d'échantillonnage devient

$$\sigma_{\text{ech}} = \sqrt{\sum_{i \in I} \left(\frac{w_i}{\sum_{j \in I} w_j} \right)^2 \sigma_{\text{ech},i}^2}.$$

4.2 Dispersion empirique

La dispersion des sondages entre eux est mesurée par

$$\sigma_{\text{disp}} = \sqrt{\frac{\sum_{i \in I} w_i (x_i - \bar{x})^2}{\sum_{i \in I} w_i}}.$$

Elle augmente lorsque des enquêtes comparables donnent des scores sensiblement différents.

4.3 Plancher d'incertitude

Le modèle ajoute un plancher

$$\sigma_{\text{res}} = 2.5.$$

Ce terme empêche une précision artificielle. Il couvre les limites non directement observées : effets institut, formulation des hypothèses, participation, erreurs de mesure, mouvements de campagne.

4.4 Écart-type nowcast

L'incertitude courante est obtenue par addition quadratique :

$$\sigma_{\text{now}} = \sqrt{\sigma_{\text{ech}}^2 + \sigma_{\text{disp}}^2 + \sigma_{\text{res}}^2}.$$

5 Projection temporelle

Pour une projection à horizon électoral, on ajoute une composante de dérive. Si Δ est le nombre de jours entre la date d'analyse et la date cible,

$$\sigma_{\text{drift}} = 1.5\sqrt{\Delta/30}.$$

L'écart-type prédictif est donc

$$\sigma_{\text{pred}} = \sqrt{\sigma_{\text{now}}^2 + \sigma_{\text{drift}}^2}.$$

Le centre $\bar{x}$ n'est pas déplacé par cette étape ; seule l'incertitude augmente.

6 Premier tour

6.1 Traitement par scénario

Les sondages de premier tour testent des listes de candidats différentes. Le modèle ne fusionne pas ces listes : chacune définit un scénario s .

Pour chaque candidat c du scénario s , on estime un score moyen $\mu_{s,c}$ et un écart-type $\sigma_{s,c}$. À la simulation k ,

$$Z_{s,c}^{(k)} \sim \mathcal{N}(\mu_{s,c}, \sigma_{s,c}^2).$$

Les scores négatifs sont tronqués :

$$\tilde{Z}_{s,c}^{(k)} = \max(0, Z_{s,c}^{(k)}).$$

Puis les scores sont renormalisés :

$$S_{s,c}^{(k)} = 100 \frac{\tilde{Z}_{s,c}^{(k)}}{\sum_{c' \in s} \tilde{Z}_{s,c'}^{(k)}}.$$

6.2 Probabilités de rang

La probabilité de terminer premier dans le scénario s est

$$\hat{P}_{s,c}^{(1)} = \frac{1}{100\,000} \sum_{k=1}^{100\,000} \mathbf{1}\{c = \operatorname{argmax}_{c' \in s} S_{s,c'}^{(k)}\}.$$

La probabilité d'être dans les deux premiers est

$$\hat{P}_{s,c}^{(2)} = \frac{1}{100\,000} \sum_{k=1}^{100\,000} \mathbf{1}\{c \in \operatorname{Top2}(S_s^{(k)})\}.$$

Lecture du Top 2 agrégé

Lorsqu'une valeur Top 2 est agrégée par candidat sur plusieurs scénarios, elle est conditionnée aux scénarios dans lesquels ce candidat est présent. Elle n'a donc pas vocation à sommer à 200 % sur l'ensemble des candidats affichés.

7 Second tour

Pour un duel d , le modèle estime un score central μ_d et un écart-type σ_d . Chaque simulation tire

$$Y_d^{(k)} \sim \mathcal{N}(\mu_d, \sigma_d^2).$$

La probabilité de victoire du candidat de référence est estimée par

$$\hat{P}_d = \frac{1}{100\,000} \sum_{k=1}^{100\,000} \mathbf{1}\{Y_d^{(k)} > 50\}.$$

Les intervalles publiés sont des quantiles empiriques de la distribution simulée. L'intervalle central à 80 % correspond aux quantiles 10 % et 90 %.

8 Pondération des scénarios

8.1 Événements structurés

Les scénarios de candidatures sont pondérés par des signaux politiques structurés : déclaration de candidature, désignation partisane, obstacle juridique, retrait, investiture, ou autre signal politiquement interprétable.

Un modèle de langage à faible température peut être utilisé pour convertir des notes factuelles en événements structurés. Cette étape ne produit pas les probabilités électorales ; elle prépare une information textuelle sous une forme exploitable par le calcul déterministe.

8.2 Score de plausibilité

Chaque événement e peut attribuer à un candidat c un impact $\delta_{e,c} \in [-5, 5]$ avec une confiance $q_{e,c} \in [0, 1]$. On pose

$$A_c = \sum_e \delta_{e,c} q_{e,c}.$$

Le score direct d'un scénario s est

$$R_s^{\text{direct}} = \sum_{c \in s} A_c.$$

Des ajustements déterministes tiennent compte des rivalités internes, de l'absence de candidats fortement signalés et de la robustesse empirique. Le score final s'écrit

$$R_s = R_s^{\text{direct}} + R_s^{\text{ajustements}} + 0.15 m_s,$$

où m_s est le plus grand nombre d'observations disponible pour un candidat du scénario.

8.3 Softmax

Les scores sont transformés en poids par

$$\pi_s = \frac{\exp(R_s/\tau)}{\sum_{s'} \exp(R_{s'}/\tau)}, \quad \tau = 2.2.$$

Poids de scénario

Le poids π_s est une mesure relative de plausibilité. Il ne prétend pas être une probabilité officielle de candidature ; il sert à pondérer les trajectoires simulées.

9 Probabilité totale de victoire

Soit s un scénario de poids π_s . Une simulation de premier tour produit deux finalistes (a, b) . Lorsque le duel est disponible, on note $P(c \mid a, b)$ la probabilité que c gagne ce duel.

La contribution brute du candidat c est

$$J_c = \sum_s \pi_s \mathbb{E}_{1^{\text{er}} \text{ tour} | s} [\mathbf{1}\{(a, b) \text{ modélisé}\} P(c \mid a, b)].$$

La couverture modélisée est

$$C = \sum_s \pi_s \mathbb{P}_{1^{\text{er}} \text{ tour} | s} ((a, b) \text{ modélisé}).$$

La probabilité conditionnelle publiée est

$$P_c^{\text{cond}} = \frac{J_c}{C}.$$

Cette normalisation signifie que la probabilité totale est lue conditionnellement aux trajectoires pour lesquelles un duel de second tour exploitable existe.

10 Précision numérique

Avec $N = 100\,000$ simulations, l'erreur numérique Monte Carlo d'une probabilité $\hat{p}$ est approximativement

$$\sqrt{\frac{\hat{p}(1 - \hat{p})}{N}}.$$

Le cas le plus défavorable est $\hat{p} = 1/2$, soit

$$\sqrt{\frac{0.25}{100000}} \approx 0.00158.$$

Cela correspond à environ 0.16 point de pourcentage. Cette incertitude numérique est inférieure aux incertitudes politiques et statistiques du modèle.

11 Limites

Limites principales

Le modèle ne corrige pas systématiquement les effets institut. Les tirages de premier tour supposent des erreurs gaussiennes indépendantes avant renormalisation, ce qui simplifie les corrélations réelles entre candidats. La pondération des scénarios dépend d'un barème de signaux politiques. Les retraits, alliances, chocs de campagne et changements de participation ne sont pas modélisés comme processus autonomes.

Ces limites ne rendent pas la méthode arbitraire ; elles définissent son domaine de validité. Le modèle est un instrument de comparaison probabiliste conditionnelle, non une garantie de résultat.