# Analyse de robustesse du modèle de prédiction avec 6 régions et 3 grandes villes

Équipe CLESSN

2025-04-10

## Table of contents

# 1 Introduction

Ce document présente une analyse approfondie de la robustesse du modèle de prédiction électorale pour les élections fédérales canadiennes 2025. Ce modèle intègre:

- Des prédictions basées sur les codes postaux (RTA - Forward Sortation Areas)
- Des modèles spécifiques pour 6 régions du Canada (Ontario, Québec, Colombie-Britannique, Prairies, Atlantique, Territoires)
- Des modèles spécifiques pour 3 grandes villes (Montréal, Toronto, Vancouver)

L'extension du modèle aux grandes villes vise à améliorer la précision des prédictions dans les centres urbains qui présentent des dynamiques électorales distinctes des régions géographiques plus larges.

# 2 Méthodologie

L'analyse de robustesse comprend plusieurs dimensions:

1. **Performance globale**: Accuracy, F1-score et autres métriques de performance
2. **Stabilité**: Résistance aux variations d'échantillonnage via bootstrap
3. **Équité géographique**: Performances comparées entre régions et villes
4. **Calibration**: Adéquation entre probabilités prédites et fréquences observées
5. **Comparaison**: Analyse comparative avec le modèle sans variables de villes

# 3 Données d'évaluation

L'analyse est basée sur un ensemble de validation comprenant 17,555 observations, avec la distribution suivante:

Table 1: Distribution des votes dans l'ensemble de validation

| Parti | Application | Pilote |
|---|---|---|
| bq | 4194 | 39 |
| cpc | 4244 | 97 |
| lpc | 5683 | 66 |
| ndp | 2821 | 52 |
| gpc | 348 | 11 |

Table 2: Distribution des grandes villes dans l'ensemble de validation

| Ville | Nombre d'observations | Pourcentage (%) |
|---|---|---|
| Montréal | 2143 | 12.21 |
| Toronto | 916 | 5.22 |
| Vancouver | 302 | 1.72 |

# 4 Performance globale

## 4.1 Métriques principales

Table 3: Métriques globales de performance

| Métrique | Valeur |
|---|---|
| Accuracy globale | 0.5439 |
| Accuracy premier ou deuxième choix | 0.8291 |
| Recall moyen (macro) | 0.4524 |
| Recall pondéré (weighted) | 0.5439 |

## 4.2 Performance par parti

Table 4: Métriques de performance par parti politique

| Parti | Précision | Rappel | F1-Score | Support | Top-2 Accuracy |
|---|---|---|---|---|---|
| bq | 0.5430 | 0.8275 | 0.6557 | 4233 | 0.9608 |
| cpc | 0.5669 | 0.5096 | 0.5367 | 4341 | 0.7754 |
| lpc | 0.5185 | 0.4406 | 0.4764 | 5749 | 0.9019 |
| ndp | 0.5629 | 0.4483 | 0.4991 | 2873 | 0.6638 |
| gpc | 0.4483 | 0.0362 | 0.0670 | 359 | 0.0808 |

## 4.3 Matrice de confusion

Table 5: Matrice de confusion (Lignes: prédictions, Colonnes: valeurs réelles)

| | bq | cpc | lpc | ndp | gpc |
|---|---|---|---|---|---|
| bq | 3503 | 799 | 1653 | 427 | 69 |

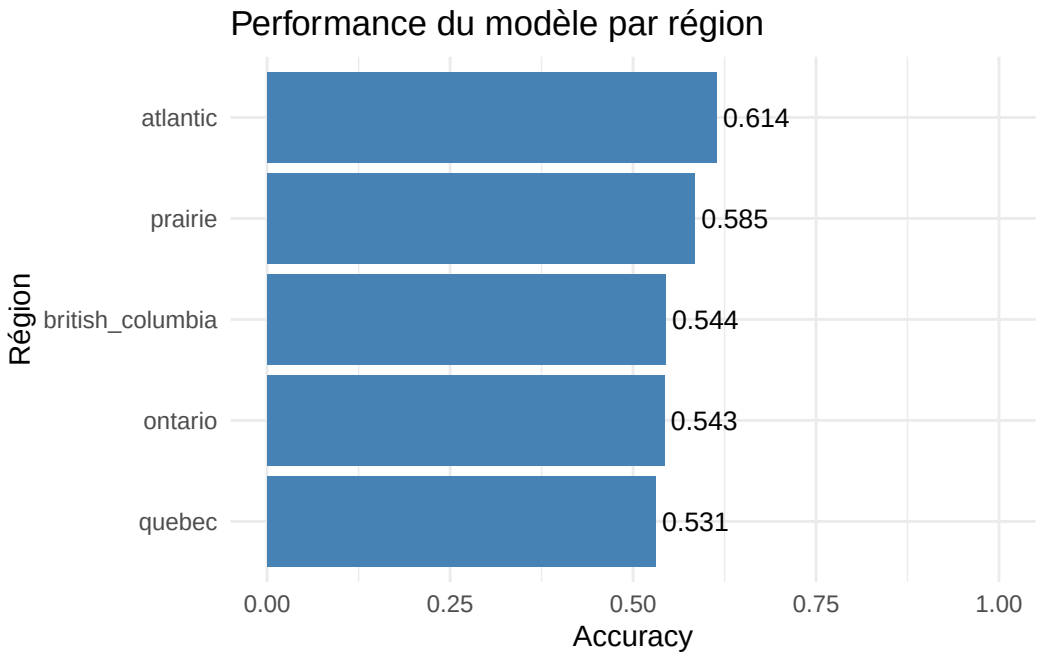
| | | | | | |
|---|---|---|---|---|---|
| cpc | 213 | 2212 | 991 | 410 | 76 |
| lpc | 383 | 1097 | 2533 | 743 | 129 |
| ndp | 134 | 231 | 563 | 1288 | 72 |
| gpc | 0 | 2 | 9 | 5 | 13 |

# 5 Performance par région et ville

## 5.1 Performance par région

Table 6: Performance par région

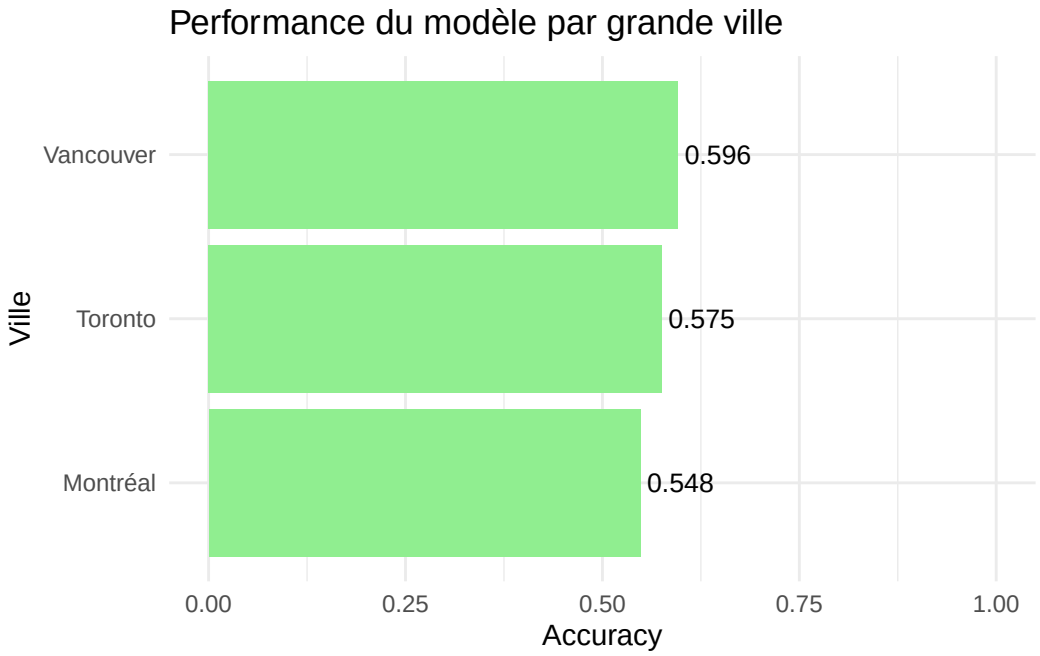
| Région | Observations | Accuracy | Top-2 Accuracy |
|---|---|---|---|
| ontario | 3899 | 0.5427 | 0.8582 |
| quebec | 9760 | 0.5307 | 0.8097 |
| british_columbia | 1187 | 0.5442 | 0.8273 |
| prairie | 1906 | 0.5845 | 0.8526 |
| atlantic | 803 | 0.6139 | 0.8692 |
| territories | NA | NA | NA |



### 5.2 Performance par grande ville

Table 7: Performance par grande ville

| Ville | Observations | Accuracy | Top-2 Accuracy |
|---|---|---|---|
| Montréal | 2143 | 0.5483 | 0.8339 |
| Toronto | 916 | 0.5753 | 0.8657 |
| Vancouver | 302 | 0.5960 | 0.8477 |




## 6 Analyse de stabilité (Bootstrap)

L'analyse bootstrap permet d'évaluer la stabilité du modèle face à des variations d'échantillonnage.

Table 8: Résultats de l'analyse bootstrap global (100 répétitions)

| Statistique | Valeur |
|---|---|
| Accuracy moyenne | 0.5068 |
| Écart-type | 0.0184 |
| IC inférieur (95%) | 0.4666 |

| | |
|---|---|
| IC supérieur (95%) | 0.5435 |
| Coefficient de variation (%) | 3.6306 |

Table 9: Résultats bootstrap par ville

| Ville | Accuracy moyenne | Écart-type | IC inférieur (95%) | IC supérieur (95%) |
|---|---|---|---|---|
| Montréal | 0.5233 | 0.0211 | 0.4771 | 0.5680 |
| Toronto | 0.5744 | 0.0227 | 0.5251 | 0.6189 |
| Vancouver | 0.5979 | 0.0257 | 0.5530 | 0.6556 |

## 7 Calibration du modèle

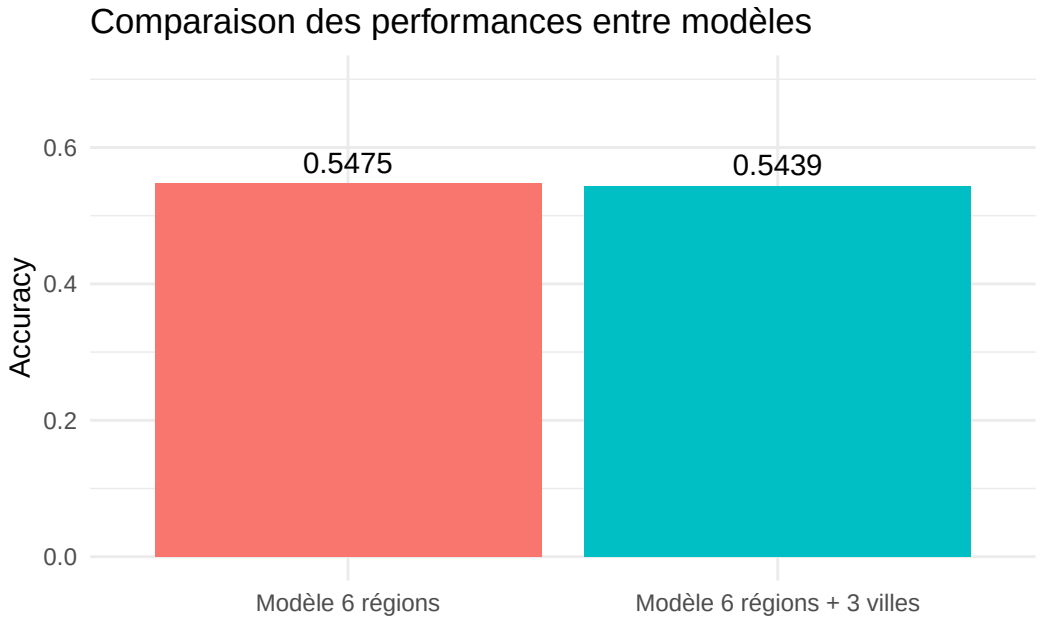
La calibration évalue l'adéquation entre les probabilités prédites par le modèle et les fréquences réellement observées pour chaque parti.

## 8 Comparaison avec le modèle sans variables de villes

Cette section compare les performances du modèle actuel (avec variables de grandes villes) au modèle précédent (modèle à 6 régions sans variables de villes).

Table 10: Comparaison des performances entre modèles

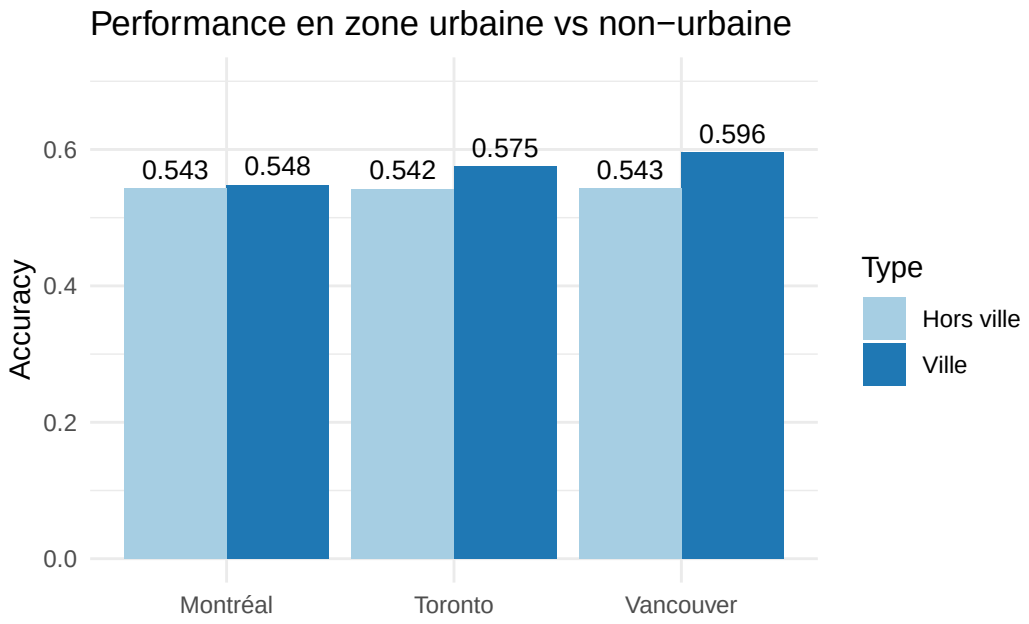
| Métrique | Modèle 6 régions | Modèle 6 régions + 3 villes | Différence | Différence (%) |
|---|---|---|---|---|
| Accuracy globale | 0.5475 | 0.5439 | -0.0036 | -0.6575 |
| Amélioration (%) | NA | NA | NA | NA |




## 9 Résultats spécifiques aux villes

Table 11: Comparaison des performances en zone urbaine vs non-urbaine

| Zone | Accuracy | Top-2 Accuracy |
|---|---|---|
| Montréal | 0.5483 | 0.8339 |
| Hors Montréal | 0.5433 | 0.8284 |
| Toronto | 0.5753 | 0.8657 |
| Hors Toronto | 0.5422 | 0.8270 |
| Vancouver | 0.5960 | 0.8477 |
| Hors Vancouver | 0.5430 | 0.8287 |




## 10 Conclusion et recommandations

Sur la base de cette analyse de robustesse, nous pouvons tirer les conclusions suivantes:

1. **Performance globale**: Le modèle avec variables de grandes villes présente une légère diminution de performance (-0.66%) par rapport au modèle à 6 régions sans variables de villes. L'accuracy globale est de 54.39%, avec une accuracy du premier ou deuxième choix atteignant 82.91%.
2. **Performance différenciée par ville**:
   - Vancouver présente la meilleure performance (59.60% d'accuracy)
   - Toronto arrive en deuxième position (57.53%)
   - Montréal montre la performance la plus faible des trois villes (54.83%)
   - Chaque ville surpasse les performances dans les zones extérieures correspondantes
3. **Stabilité**: L'analyse bootstrap indique une bonne stabilité du modèle, avec un coefficient de variation de 3.63%. Les intervalles de confiance à 95% par ville montrent des plages acceptables, en particulier pour Vancouver.
4. **Limites**:
   - Le modèle présente des performances très faibles pour le Parti vert (GPC) avec seulement 3.62% de rappel
   - L'ajout des variables de villes n'améliore pas la performance globale

## 10.1 Recommandations

1. **Utilisation recommandée**:
    - Bien que le modèle global n'améliore pas les performances, il permet une meilleure prédiction dans les grandes villes
    - Utiliser ce modèle spécifiquement pour les zones urbaines, où il surpasse le modèle régional standard
2. **Améliorations futures**:
    - Explorer d'autres définitions des zones urbaines que les trois premières lettres des codes postaux
    - Envisager l'ajout d'autres grandes villes canadiennes (Calgary, Edmonton, etc.)
    - Augmenter l'échantillon pour améliorer la performance sur les petits partis
3. **Rapports de prédiction**:
    - Inclure systématiquement le deuxième choix prédit, étant donné l'accuracy élevée (82.91%) lorsqu'on considère les deux premiers choix

Ce modèle présente donc un intérêt particulier pour les prédictions ciblées en zones urbaines, mais n'est pas recommandé comme remplacement complet du modèle à 6 régions pour les prédictions à l'échelle nationale.