Damien Dallon & Nathanaël Lefèvre
M2 IA CMI Informatique

Decision Support

Application BI

Introduction

Dans le cadre de ce projet d'application BI nous avons choisi d'aborder la thématique de la théorie du choix social. Il s'agit d'implémenter différentes méthodes vues en cours et les appliquer sur des jeux de données fournis afin d'en comparer les résultats.

Données

Structure des données

Nous disposons de 3 fichiers représentant des profils de votes. Chaque profil présente une colonne par électeur, les éléments d'une colonne correspondant à des numéros de candidats classés par ordre de préférence.
Le premier profil contient 1000 électeurs et 9 candidats numérotés de 1 à 9. Les deux autres profils contiennent chacun 10 000 électeurs et 12 candidats numérotés de 1 à 12.

Analyse préliminaire des votes

Voici ci-dessous pour chaque profil de vote une simple analyse de la distribution des votes par candidat : à gauche le nombre de votes par candidat pour les 1ère, 2nd et 3ème places ; à droite la moyenne du rang par candidat indiquant la tendance du candidat à être apprécié par les électeurs.

Notons que la moyenne du rang correspond en réalité au même principe que la méthode de Borda. En effet, Pour calculer le score de Borda, on procède ainsi :

Soit C l'ensemble des candidats
Soit E l'ensemble des électeurs
Soit $V = [v_{i,j}], (i,j) \in [1,|E|] * [1,|C|]$ la matrice des classements
Soit $S = [s_j], k \in [1,|C|]$ le score de borda pour les candidats
On note ind la fonction indicatrice
Alors le score de Borda se calcule ainsi :

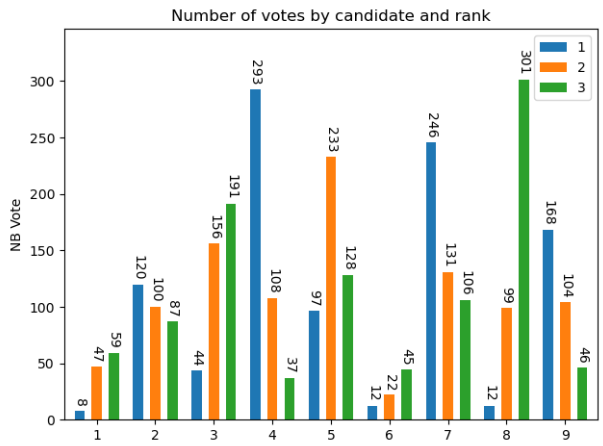
$$S_k = \sum_{j \in [1,|C|]} (|C| - j + 1) \sum_{i \in [1,|E|]} ind\{V_{i,j} == k\}$$

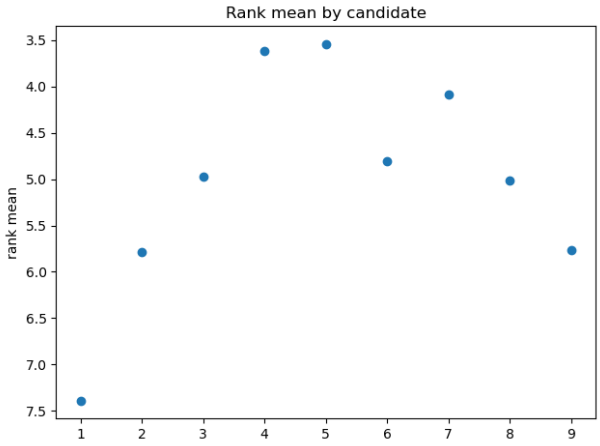
Or notre rang moyen se calcule comme:

$$R_k = \frac{\sum_{j \in [1,|C|]} (|C| - j + 1) \sum_{i \in [1,|E|]} ind\{V_{i,j} == k\}]}{|E|}$$

D'où: $R_k = \frac{S_k}{|E|}$

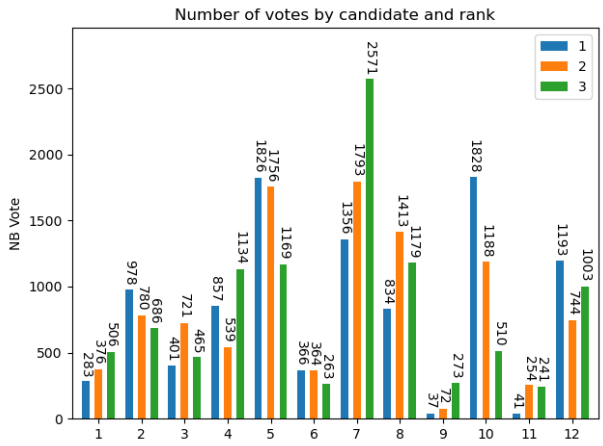
Profil 1 -

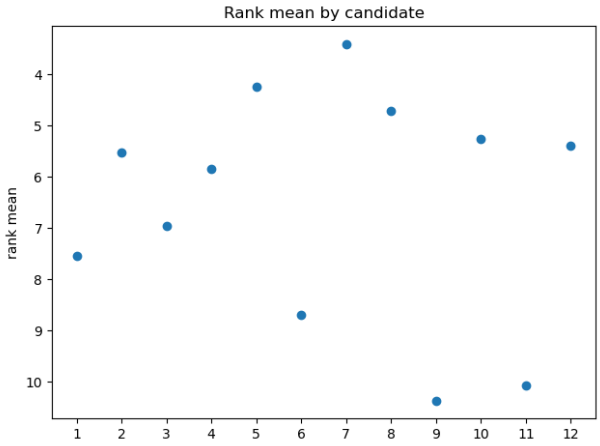




Pour le profil 1, on remarque que le candidat 4 a le plus de première place parmi les candidats et son rang moyen dans les votes est très élevé : il semble donc s'agir d'un candidat prometteur.
Cependant, on remarque que le candidat 5 a un rang moyen plus important et, bien qu'il ait peu de premières places, il est en tête sur le nombre de seconde place.
Par ailleurs, bien que le candidat 7 ait beaucoup de premières places, il est 3ème en termes de rang moyen.

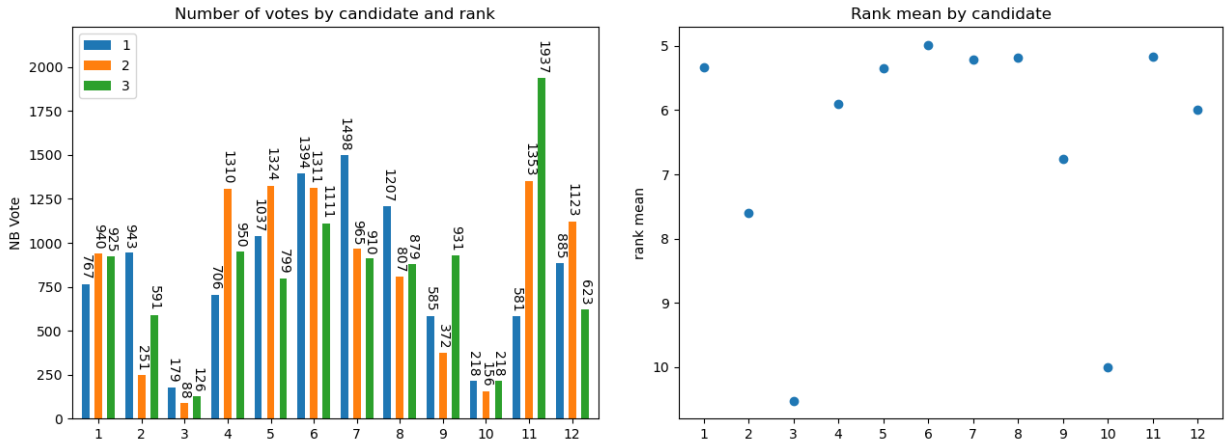
Profil 2 -





Pour le profil 2, le candidat 7 ressort avec la meilleure moyenne de rang et un beaucoup de premières mais surtout plus de secondes et troisièmes places que tout autre candidat.
Le candidat 5 est également un bon élément puisqu'il est tout juste second en termes de rang moyen et de nombre de premières places. De plus, il a un nombre de seconde place presque aussi important que son nombre de premières places.
Enfin, le candidat 10 apparaît avec le plus grand nombre de premières places mais il est 4ème en termes de rang moyen.

Profil 3 -



Pour le profil 3, le candidat 7 ressort avec le plus de premières places et le candidat 6 avec le meilleur rang moyen. Les candidats 1, 4, 5, 6, 7, 8 et 11 ont en grande partie un rang moyen assez similaire et se distinguent surtout sur le nombre de premières et secondes places. Le candidat 6, encore une fois, propose à la fois un nombre élevé de premières et secondes places.

Méthodes de votes

Nous avons implémenté les méthodes suivantes :

- Votes à n-tours, en particulier à 1 tour, à 2 tours et vote alternatif (celui qui a moins de premières places est éliminé à chaque tour)
- Borda
- Condorcet

Résultats des méthodes de votes

Voici les résultats des gagnants obtenus pour les trois profils :

	Profil 1	Profil 2	Profil 3
Vote à 1 tour	4	10	7
Vote à 2 tours	4	5	7
Vote alternatif	4	7	6
Borda	5	7	6
Condorcet	5	7	Pas de gagnant

Conclusions sur les résultats

En comparant avec nos observations sur les graphiques précédents pour chaque profil, nous observons que les résultats sont plutôt cohérents.

Nous pouvons décrire les observation suivantes :

- Lors d'un vote à un tour, le gagnant est celui qui obtient le plus de premières places. En effet c'en est la définition même, cependant cela est bien observable dans les trois profils.

- Lors d'un vote à deux tours, le gagnant est généralement celui qui a un meilleur taux de votes pour les premières et deuxièmes places. Cela est bien observable pour le profil 2 dans lequel les candidats 10 et 5 ont quelques votes de différences pour la première place, mais le 5 en a sensiblement plus pour la deuxième place.
  Au contraire, le profil 3 contredit cette observation car le candidat 6 devrait gagner sur le 7. Cependant on peut supposer qu'un certain nombre de votes pour la deuxième place donnés au candidat 6 proviennent des électeurs préférant le candidat 7, ainsi ces votes ne vont pas effectivement au candidat 6.

- Le gagnant selon Borda, comme expliqué précédemment, est celui qui obtient la meilleure place en considérant la moyenne des préférences des électeurs. On peut considérer que la méthode Borda est celle qui offre le "meilleur compromis" pour élire un candidat et contenter le plus grand nombre d'électeurs.

- La méthode de Condorcet désigne le candidat qui bat en duel tous les autres candidats (selon le nombre de voix évidemment, sinon cela ne serait plus un vote mais un battle royal :)). Ainsi s'il y a un vainqueur ce sera forcément le même que selon la méthode de

Borda. Cependant comme l'indique le profil 3 le vainqueur selon Borda n'est pas forcément un vainqueur de Condorcet, en particulier lorsque plusieurs des meilleurs candidats ont un rang moyen similaire, ce qui apporte une plus grande probabilité qu'aucun candidat ne batte tous les autres.

- Le vote alternatif n'est pas le mieux représenté parmi les résultats présentés ci-dessus. Malgré cela nous pouvons supposer que le gagnant est souvent similaire à Borda car le candidat ayant le moins de premières places est éliminé tour à tour et les votes de ses électeurs se répartissent pour les candidats restants. Ainsi ce procédé peut se rapprocher d'un moyenne, en particulier si le nombre de candidats est important (ce qui est le cas des profils 2 et 3 par rapport au profil 1).

Analyse de robustesse

Afin d'étudier la robustesse des méthodes de votes, nous avons décidé de perturber les votes et de faire varier la taille de l'échantillon d'électeurs. Nous avons utilisé la méthode bootstrap afin de tirer 10 échantillons de votes avec remise en suivant la distribution des votes de départ. Une analyse préliminaire analogue à la partie "Données / Analyse des votes" de ce rapport a été effectuée sur les tendances moyennes des échantillons obtenus (la **moyenne** du rang moyen des candidats et la **moyenne** du nombre de places par candidats et par rang).

La méthode a été appliquée pour une taille d'échantillon allant de 1% à 100% de la taille totale de l'échantillon d'origine. On a donc 100 étapes avec 10 échantillons tirés à chaque étape pour un total de 1000 échantillons.

Par exemple, pour 1000 électeurs de départ, on a fait :

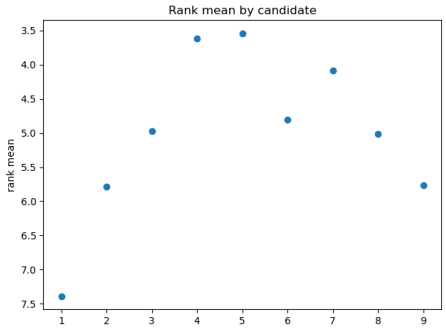
- 10 tirs aléatoires de 100 votes
- 10 tirs aléatoires de 200 votes
- …
- 10 tirs aléatoires de 1000 votes

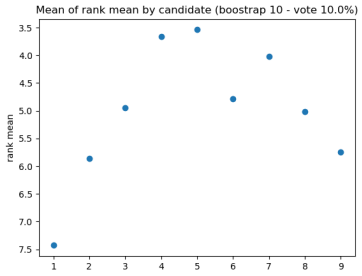
Ainsi, nous pouvons voir en fonction du nombre de votes:

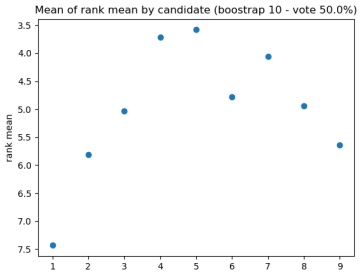
- Les résultats des différentes méthodes sur les 10 échantillon (indique la robustesse)
- L'écart-type entre les rang moyens de chaque candidat pour les 10 échantillons (indique la dispersion des donnée : le degré de non ressemblance à la distribution originale)
- La moyenne du rang pour chaque candidat (Indique naturellement les résultat de Borda)

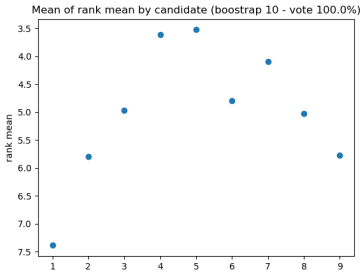
Profil 1

Référence (échantillon d'origine)



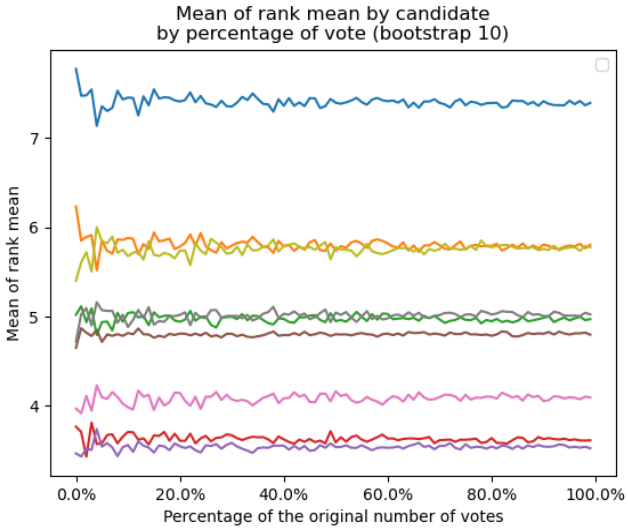




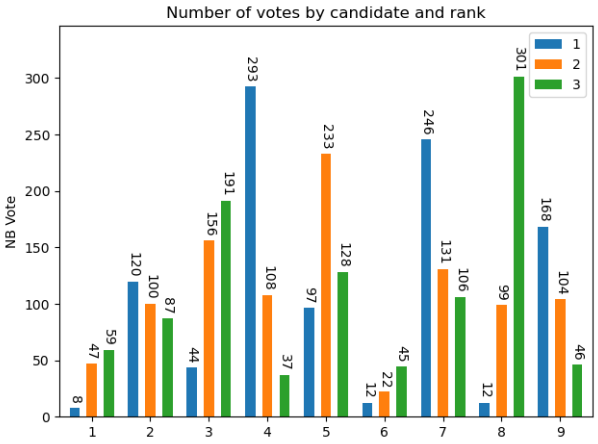


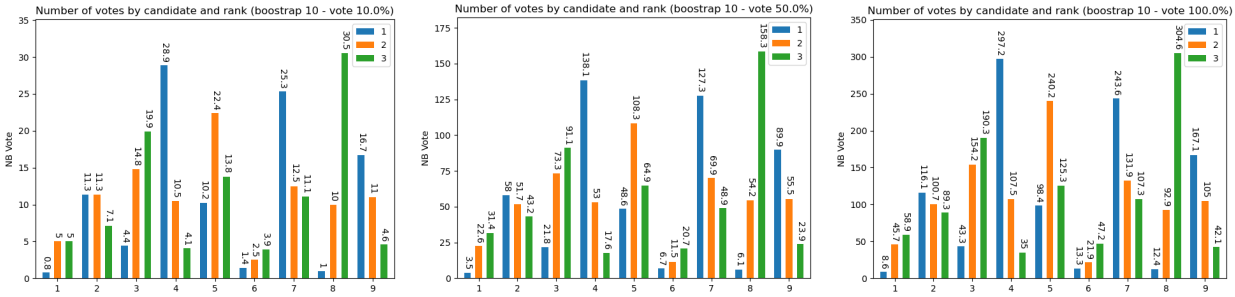
En faisant la **moyenne** du **rang moyen** de chaque candidats sur les 10 échantillons, pour une taille d'échantillon de 10% (du nombre d'électeur originel), de 50% ou 100%, le rang moyen bootstrapé est en moyenne très similaire au rang moyen d'origine (Le classement est le même).

On peut donc penser que la méthode de Borda ne serait pas affectée. Pourtant, les résultats qui suivent prouvent le contraire. En effet, en affichant cette fois-ci la moyenne du rang moyen par candidat en fonction de la taille de l'échantillon, on voit que pour certains candidats dont les rang moyen sont proches, ces candidats sont tantôt l'un devant l'autre, tantôt l'inverse :



Référence (échantillon d'origine)

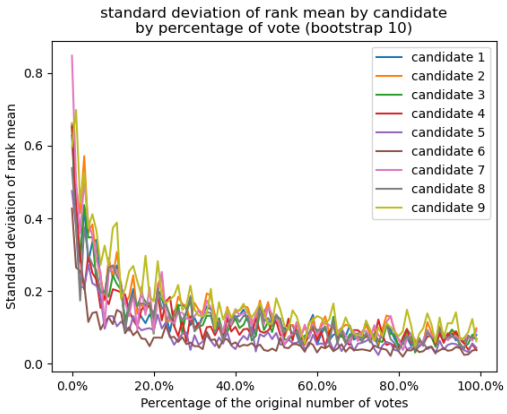




Pour ce qui est de la distribution des 1ère, 2nd et 3ème places, avec une taille de 10, 50 et 100% de la taille de l'échantillon d'origine, on a plus ou moins la même distribution. Pour 10% on voit néanmoins que le candidat 2 a en moyenne le même nombre de 1ère et 2nd places sur les 10 échantillons tandis qu'à 50% et 100%, il a moins de 2nd places que de 1ère.

Nous pourrions faire un graphique du nombre de 1ère, 2nd et 3ème places de chaque candidat ce en fonction du nombre d'électeurs mais nous n'allons pas le faire pour la simple raison scientifique qu'il ne faut pas abuser. Nous garderons simplement en tête qu'**avec un échantillon de petite taille, le profil de vote ressemble moins au profil de vote d'origine**.

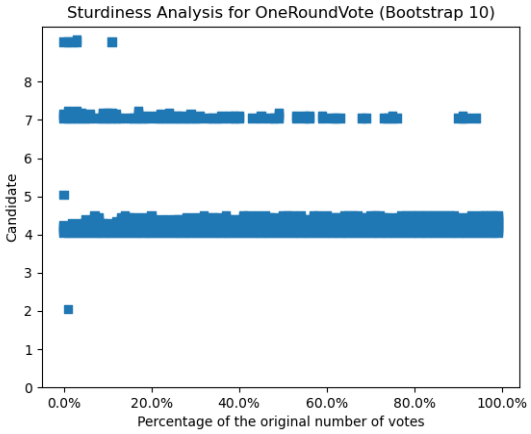
On peut en effet voir que l'écart type entre les moyenne de rank des candidats d'un échantillon à l'autre décroît de façon logarithmique. Ainsi, plus la taille d'échantillon augmente, plus les échantillons sont semblables mais cette tendance va en ralentissant :

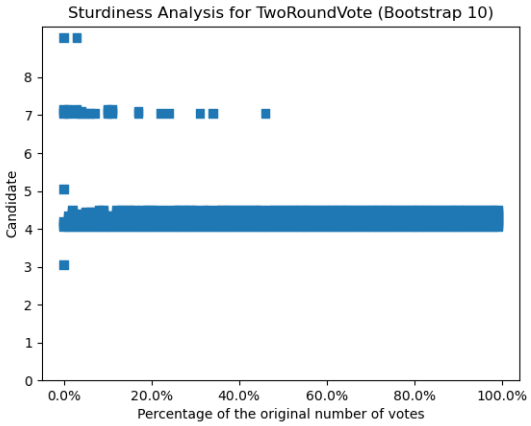


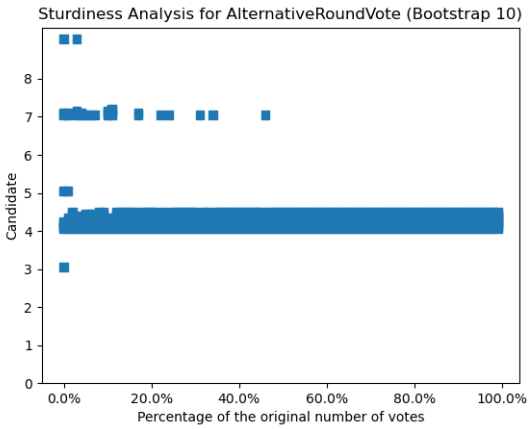
Voyons le résultat des méthodes:

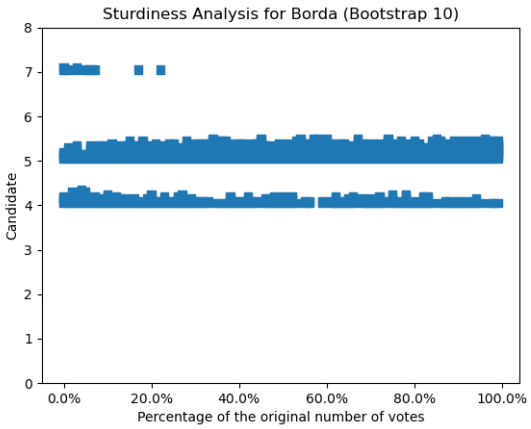
Pour chaque graphique, on voit pour chaque taille d'échantillon le résultat de la méthode sur les 10 tirs aléatoires. Par exemple pour le premier graphique, on voit que sur 10 échantillons aléatoires de très petite taille, OneRoundVote (Le vote à un tour) a souvent élu 4 ou 7, et un peu 5 et 2. Sur des échantillons de taille très proche de l'échantillon original, il prédit toujours 4 mais sur des tailles un peu inférieures il prédit aussi 7.

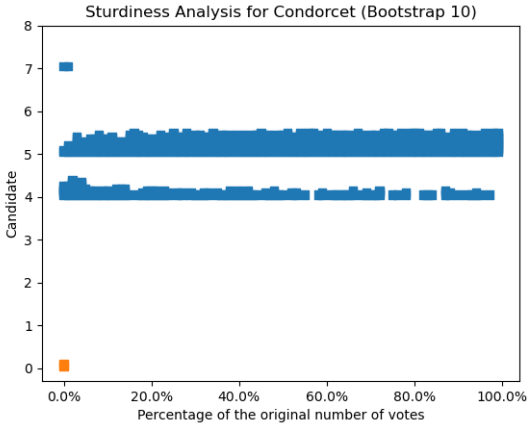
Les cases oranges pour le candidat 0 indique qu'aucun vainqueur n'a pu être déterminé.











Analysons la robustesse par type de vote:

- Vote à un tour (OneRoundVote) : Le vote à un tour n'est pas très robuste. Si on perturbe un peu les votes et que l'on a un peu moins d'électeurs qu'à l'origine, on voit que la méthode peut élire 7 de temps en temps. Notons que 7 ressortait aussi comme un bon candidat dans notre analyse préliminaire. Cette méthode est trop "hésitante".

- Vote à deux tours (TwoRoundVote) : Cette méthode est très robuste. Si pour une taille d'échantillon très faible la méthode peut élire des candidats très différents, il suffit d'un nombre d'électeur supérieur à 50% de la taille d'origine pour avoir le même résultat qu'avec le vrai échantillon, malgré le fait que les votes soient perturbés ! (rappelons que les votent suivent la distribution d'origine grâce à un tirage au sort **avec remise**). On peut voir en effet qu'à partir de 50%, la méthode élit toujours le candidat 4. Dans une France où le poids de l'abstention se fait de plus en plus ressentir, cette méthode de vote semble intéressante pour peu que les électeurs soient représentatif de l'opinion des français. En effet, si les non-votants sont en réalité tous en faveur d'un candidat commun, les votes réels (des non-abstentionnistes) ne reflètent alors pas la vraie distribution de l'opinion française. Cette méthode peut donc sembler plutôt bien adaptée à notre démocratie. Le débat reste ouvert.

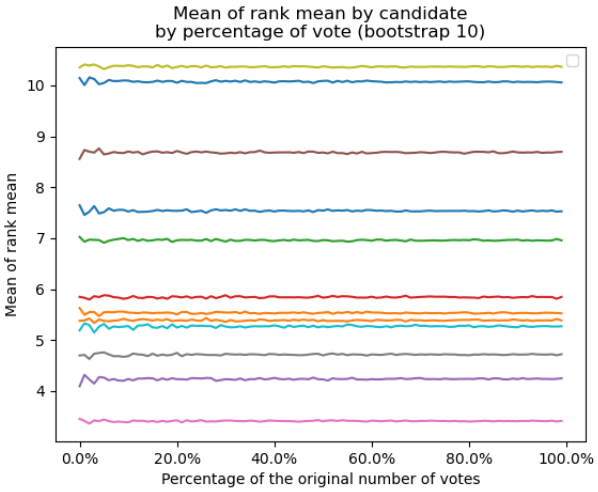
- Vote alternatif (AlternativeRoundVote) : Cette méthode est tout aussi robuste que le vote à deux tours. Les deux graphiques obtenus sont quasiment les mêmes (on peut observer de très petites différences en observant bien). Cependant, cette méthode requiert un nombre de tours potentiellement plus important que le vote à 2 tours.

- Vote de Borda (Borda) : Cette méthode est très peu robuste. Pour un échantillon de la taille de celui d'origine, la méthode élit souvent le candidat 5 comme pour le vrai échantillon, mais il lui arrive d'élire le candidat 4. Celà reste cohérent avec notre analyse préliminaire qui prédisait les candidats 4, 5 et 7 comme intéressants. Cependant, la taille d'échantillon a assez peu d'impact. En dessous de 30% de la taille d'origine, la méthode choisit parfois le candidat 7 (comme par hasard) sur certains des 10 échantillons. Mais au-delà de 30%, lorsque l'écart-type est en moyenne inférieur à environ 0.2 la méthode "hésite" entre les candidats 4 et 5 de la même manière. **Cependant**, cette méthode peut être intéressante si l'on doit **élire 2 personnes au même poste** au lieu d'une seule ! En effet, dans ce cas, la méthode s'avère être très robuste puisqu'elle choisit toujours les mêmes candidats au-delà de 30% de d'électeurs. Bien sûr, il ne faut pas avoir d'intérêt pour le classement entre les deux personnes choisies dans ce cas.

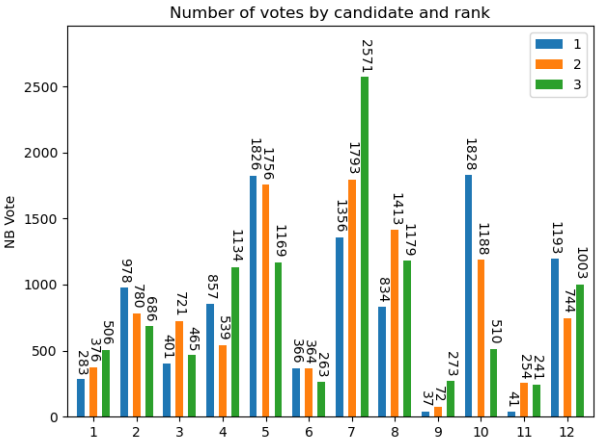
- Vote de Condorcet (Condorcet) : Cette méthode présente ici les mêmes caractéristiques que Borda, à la différence près qu'un échantillon supérieur à 5 ou 10% à peine suffit pour avoir les mêmes résultats qu'avec la taille d'échantillon d'origine. De plus, dans certains cas, il n'est pas possible pour la méthode de faire un choix.

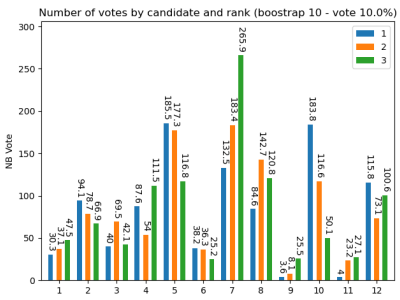
Profil 2

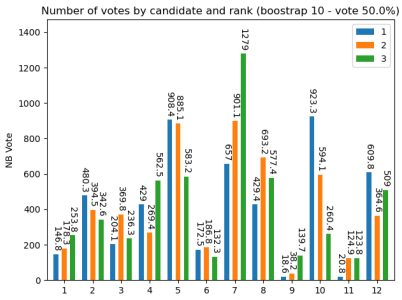
L'analyse des échantillons du profil 2 confirment les résultats présentés avec le profil 1.
Ce profil présente cependant une différence : la **moyenne du rang moyen** sur les 10 échantillons produit un classement des candidats qui **reste constant** en fonction de la taille d'échantillon.

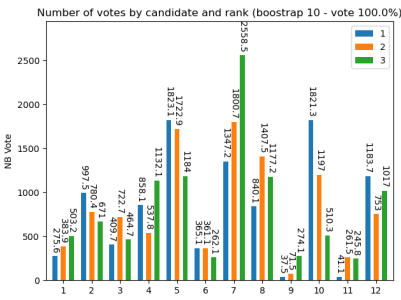


De même, à 10, 50 et 100%, on a un profil de 1ère, 2nd et 3ème place par candidat très similaire :

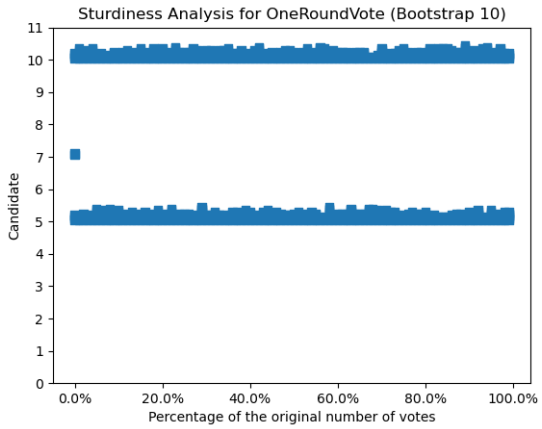


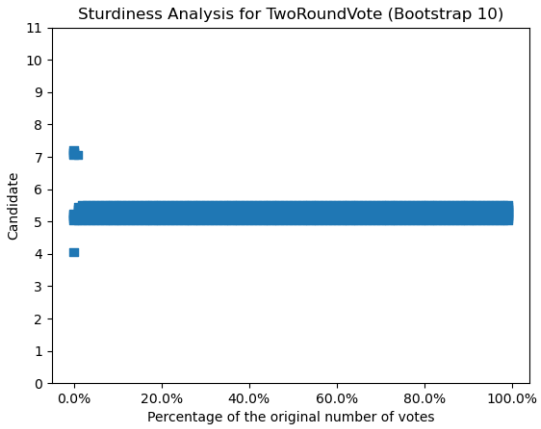


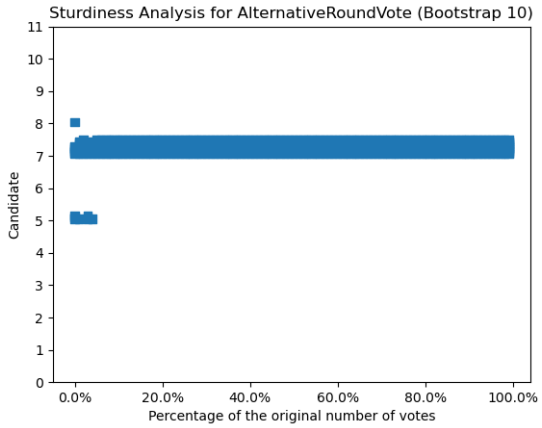


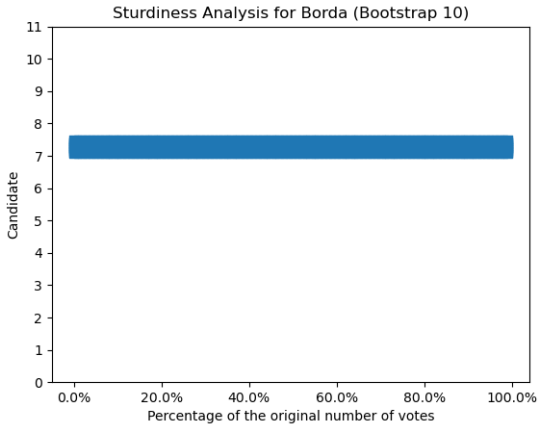


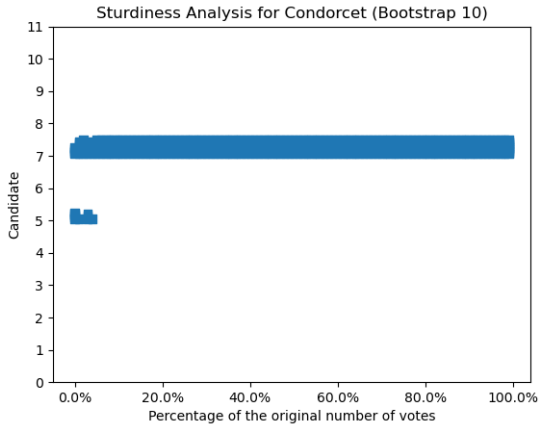
Voyons le résultat des méthodes:



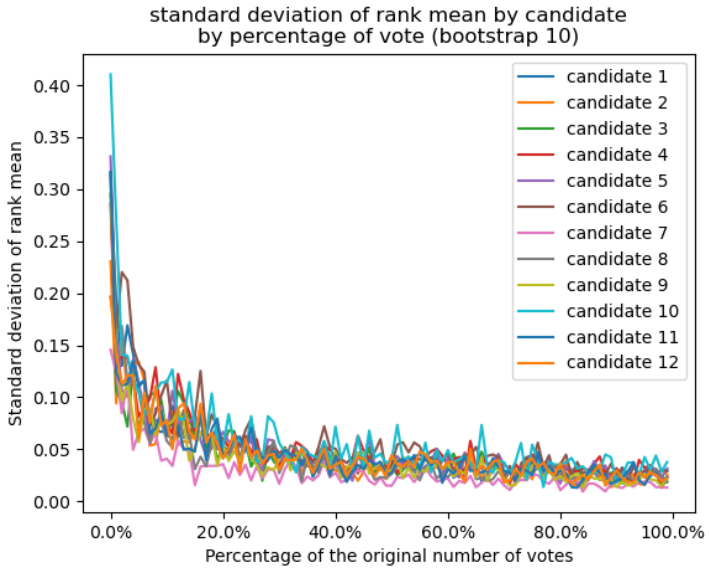








Les résultats confirment notre analyse du profil 1 pour les méthodes à 1 et 2 tours et le vote alternatif. Notons que la grande taille de l'échantillon d'origine permet de faire baisser l'écart type entre les 10 échantillons très tôt :



Ainsi, si dans le profil 1 il fallait attendre d'avoir pris 50% de la taille de l'échantillon d'origine pour descendre en dessous de 0.15, on descend ici en dessous de ce seuil au alentours de 5%. Il n'est donc pas étonnant que les résultats de nos méthodes se stabilisent au-delà de 5% d'électeurs.

Cependant, dans ce cas, Bordas et Condorcet sont très robustes.

Conclusion de l'analyse de robustesse

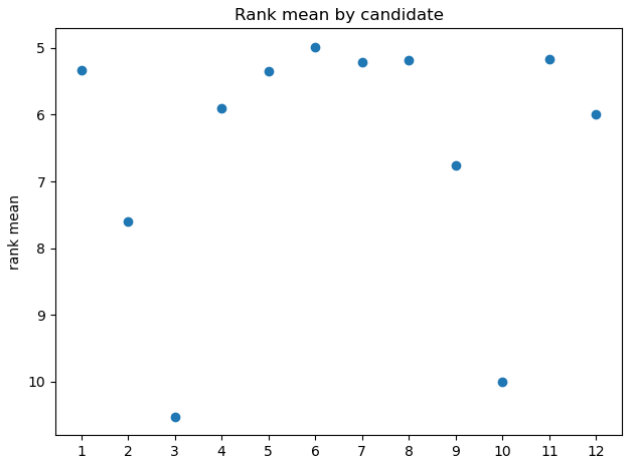
Il ressort que le vote à 2 tours semble très robuste et adapté à notre société. Le vote alternatif est comparable au vote à 2 tours mais demande potentiellement plus de temps pour un résultat similaire. Les votes de Borda et de Condorcet sont relativement robustes mais celà dépend de l'écart entre les candidats. Si deux candidats sont trop proches, ces méthodes ont du mal à en choisir un plutôt qu'un autre si on perturbe légèrement les votes. Dans certains cas, ces méthodes peuvent donc être utilisées pour élire 2 personnes plutôt qu'une mais ce n'est pas non plus garanti. Le vote à 1 tour est quant à lui très peu robuste quel que soit la taille de l'échantillon d'origine.

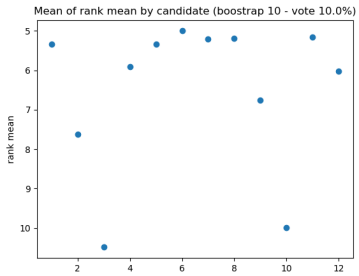
Les résultats du profil 3 sont en annexe et remettent quelque peu en cause notre analyse. Il semblerait qu'un phénomène rende compliqué l'élection d'un candidat avec la méthode de Condorcet qui n'arrive pas à statuer. Il est possible que le même phénomène explique la différence avec les résultats que nous avons évoqué ci-dessus. Dans ce cas, seule la méthode de Borda présente une bonne robustesse à partir de 50%

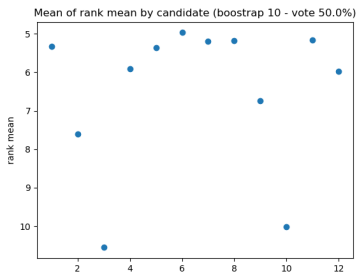
PS : Les pourcentages sont censés commencer à 10% sur les graphiques. Les échantillons de plus petite taille font en effet 10% de la taille de l'échantillon d'origine dans tous les cas. Il faut donc ajouter 10% à tous les pourcentages présentés dans l'analyse pour être tout à fait rigoureux. Celà ne change pas grand-chose et les graphiques sont très longs à générer. Pour garder en clarté, nous avons préféré ne pas corriger les pourcentages dans l'analyse et garder la cohérence entre ces derniers et les histogrammes.

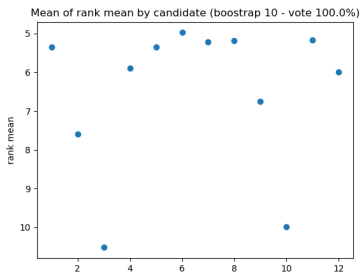
Annexes

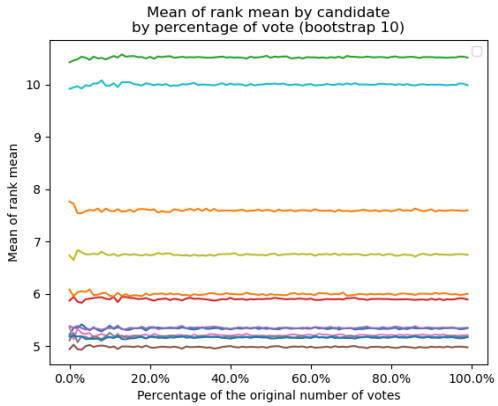
Référence (échantillon d'origine)

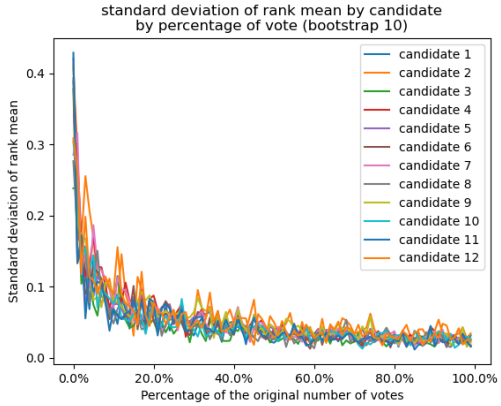




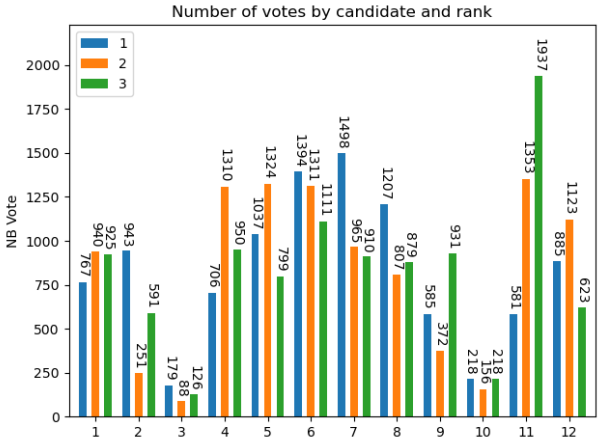


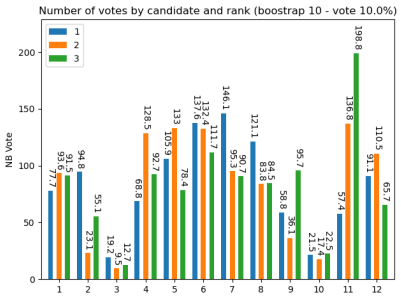


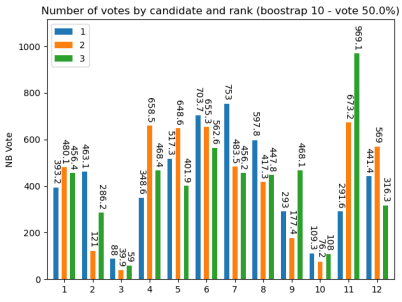


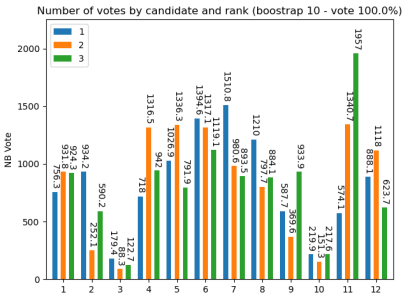


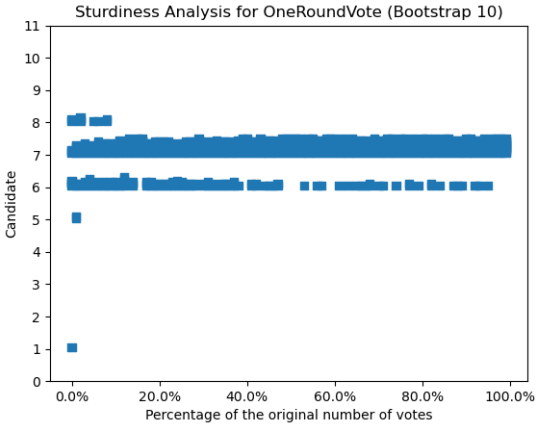
Référence (échantillon d'origine)

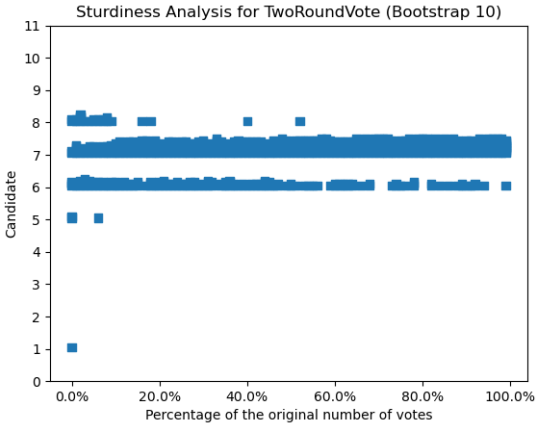


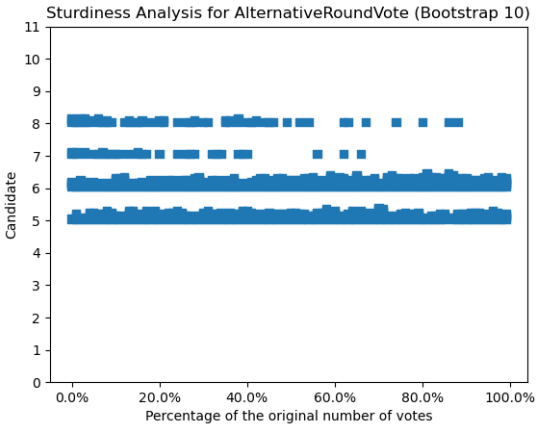


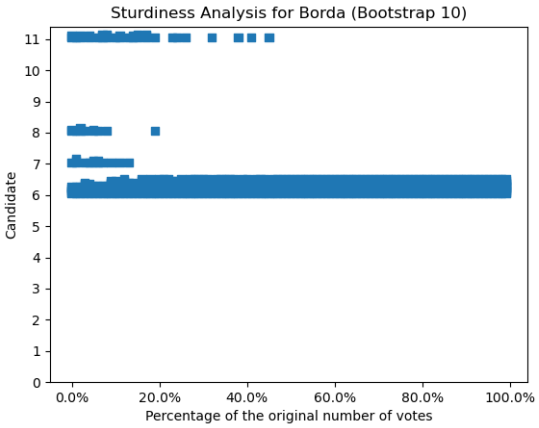


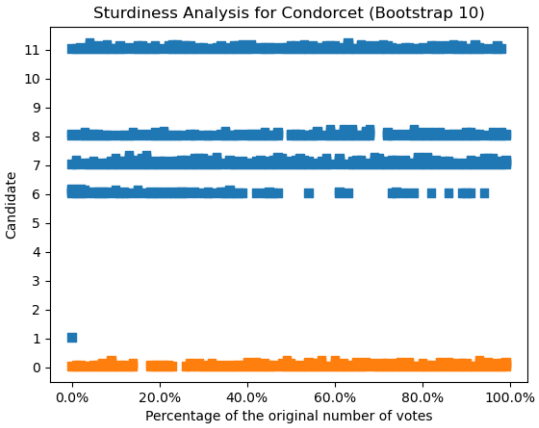



Sturdiness Analysis for OneRoundVote (Bootstrap 10)
Candidate
Percentage of the original number of votes


Sturdiness Analysis for TwoRoundVote (Bootstrap 10)
Candidate
Percentage of the original number of votes


Sturdiness Analysis for AlternativeRoundVote (Bootstrap 10)
Candidate
Percentage of the original number of votes


Sturdiness Analysis for Borda (Bootstrap 10)
Candidate
Percentage of the original number of votes


Sturdiness Analysis for Condorcet (Bootstrap 10)
Candidate
Percentage of the original number of votes