Simon Aubé (111 158 058)
Nassim El Massaudi (536 916 232)
Alexandre Moreau (111 225 281)

Analyse et traitement de données massives
GLO-7027

Rapport 1

Travail présenté à
M. Richard Khoury

Département d'informatique et de génie logiciel
Université Laval
Hiver 2022

Prédiction des intentions de vote de l'Étude Électorale Canadienne 2019: Rapport 1

Simon Aubé[‡*], Nassim El Massaudi[‡*], Alexandre Moreau[‡*]
[‡] Université Laval
[*] Ces auteurs ont contribué également

Abstract

De nombreux facteurs influencent les préférences des électeurs au cours d'une campagne électorale, et ceux-ci ne sont pas pleinement compris. La prédiction d'intentions de vote n'est donc guère aisée. De vastes ensembles de données combinant les intentions de vote à des attributs idéologiques et socio-économiques pour un grand nombre d'individus peuvent contribuer à élucider les déterminants du vote. Les résultats de l'Étude électorale canadienne de 2019 représentent justement un tel ensemble de données. Nous proposons donc de les utiliser afin d'élucider les relations qui unissent une variété d'attributs aux intentions de vote des Canadiens et, ensuite, de prédire ces dernières. Dans ce premier rapport, nous identifions d'abord plusieurs prédicteurs potentiels des intentions de vote et caractérisons les corrélations entre eux, par deux approches complémentaires. Puis, nous établissons que les arbres de décision et les ensembles en forêt aléatoire constituent des algorithmes particulièrement adaptés à ce problème d'apprentissage et établissons les grandes lignes d'un régime d'entraînement adéquat. Des travaux futurs, présentés dans un second rapport, viseront à valider l'ensemble de ces conclusions par la prédiction d'un sous-ensemble d'intentions de vote.

1. Introduction

Lors d'élections démocratiques, les choix des citoyens ont de multiples causes. Les électeurs votent évidemment selon leurs préférences et leurs valeurs, mais les positions idéologiques relatives des partis les uns par rapport aux autres [1] de même que des considérations stratégiques – visant notamment l'expression d'un vote ayant un plus grand impact [2] – entrent aussi en jeu. En conséquence, la prédiction des intentions de vote de groupes ou d'individus n'est pas nécessairement aisée. Celle-ci peut pourtant être très utile, entre autres pour les partis politiques en cours de campagne, et est donc un sujet de recherche fréquent. Les prévisions électorales se basent habituellement sur des sondages, mais ceux-ci présentent souvent des biais dus à un échantillonnage non-représentatif [3]. Le développement de nouveaux outils ne se basant pas sur l'auto-déclaration des intentions de vote – par exemple en ayant recours à des indicateurs socio-économiques [4] – est donc souhaitable. Les déterminants des intentions de vote ne sont cependant pas pleinement élucidés, ce qui complique cette tâche.

Afin de mieux comprendre les différents facteurs qui guident les choix des électeurs, l'élection fédérale canadienne de 2019 constitue un système modèle tout indiqué. Celle-ci a en effet mis en scène des partis représentant une large gamme de positions idéologiques, tandis que le système électoral canadien peut favoriser le vote stratégique. Un vaste ensemble de données sur cette campagne électorale est d'ailleurs disponible, grâce à l'Étude électorale canadienne [5]. Cette dernière contient des informations récoltées auprès de 37 822 citoyens et résidents permanents canadiens, qui concernent autant leurs intentions de vote et leur perception des partis et figures politiques que leurs positions idéologiques ainsi que leur statut socio-économique. Elle se divise en deux parties: une enquête préélectorale effectuée auprès de la totalité des répondants entre le 13 septembre et 21 octobre 2019 (jour de l'élection) ainsi qu'une enquête postélectorale réalisée du 24 octobre au 11 novembre 2019 auprès de 10337 des participants de départ. Au total, ces deux volets ont fourni 600 attributs

se rapportant aux différents répondants, bien que plusieurs soient plutôt des métadonnées sur le sondage et qu'aucun individu n'ait répondu à la totalité des questions.

Un aussi vaste ensemble de données représente une opportunité exceptionnelle d'en apprendre davantage sur les déterminants des intentions de vote au sein d'une population, et c'est justement notre objectif. Plus précisément, nous visons à identifier les attributs les plus informatifs sur les intentions de vote au moyen d'approches statistiques, puis à valider ces conclusions en prédisant avec succès un sous-ensemble d'intentions de vote qui a préalablement été masqué dans l'ensemble de données. Dans ce premier rapport, nous présentons d'abord des analyses préliminaires visant à caractériser les relations entre les différents attributs ainsi qu'à identifier ceux qui prédisent le mieux les intentions de vote. Puis, à la lumière de ces analyses, nous discutons des attributs qui seront pris en compte en priorité dans la cadre de la tentative subséquente de prédiction. Nous terminons en identifiant le type d'algorithme d'apprentissage qui semble le plus approprié pour ce problème et en décrivant la procédure d'entraînement qui sera utilisée.

2. **Méthodes**

2.1. **Préparation du jeu de données**

À première vue, le jeu de données contient beaucoup de données qui ne seront pas utiles pour un modèle de prédiction. C'est le cas par exemple des attributs qui proviennent du sondage effectué après les élections, des métadonnées et des données qui n'ont qu'une seule valeur possible. Il y a aussi des données qui ne seront pas disponibles pour les données de test à cause de leur nature trop révélatrice sur l'intention de vote. Nous avons aussi conclu que les attributs comportant du texte libre pouvaient être retirés car ils sont souvent complémentaires à une question à choix multiples (lorsqu'un individu indique "Autre" comme choix de réponse) et donc que l'information apportée ne vaudrait pas la complexité de gérer les multiples réponses possibles. Nous devions donc retirer ces données avant de procéder à l'analyse du jeu de données.

Nous avons ainsi retiré toutes les données du sondage post-élections, soit tous les attributs qui sont représentés par le préfixe "pes19". De même, nous avons aussi retiré les métadonnées, soit les attributs qui représentent les dates de début et de fin du questionnaire, l'ID de réponse, les différents *flags*, les poids, etc. Nous avons aussi supprimé les données comportant du texte libre et les attributs non disponibles pour les données de test. Enfin, nous avons retiré les attributs qui n'ont qu'une seule valeur possible. Ces attributs sont $cps19\_consent$ et $cps19\_yob\_2001\_age$. Après avoir fait ce ménage initial, l'ensemble de données comportait 234 attributs. Nous devions ensuite combiner les colonnes représentant la valeur à prédire. Ces colonnes sont $cps19\_votechoice$, $cps19\_votechoice\_pr$, $cps19\_vote\_unlikely$, $cps19\_vote\_unlike\_pr$ et $cps19\_v\_advance$. Celles-ci doivent être combinées, car elles représentent toutes le choix de vote d'un individu, selon ses réponses à d'autres questions. Ces colonnes sont donc complémentaires et leur combinaison crée le vecteur des intentions de vote composites utilisé dans le cadre des analyses suivantes.

Les tests statistiques subséquents ont ainsi été effectués sur 230 attributs.

2.2. **Dispersion et valeurs fréquentes**

Afin de trouver des données aberrantes ainsi que des attributs qui pourraient nous indiquer des informations cruciales sur les données sans méthodes mathématiques poussées, nous nous sommes penchés sur le coefficient de variation ainsi que le mode des données.

Le coefficient de variation a été calculé sur les données numériques et le mode a été calculé sur les données nominales et ordinales.

Ces statistiques ont été calculées selon chaque parti, révélant donc les différences de variation et de mode pour chacun d'entre eux.

2.3. **Corrélations avec les intentions de vote**

Afin d'évaluer la corrélation des différents attributs de l'ensemble de données, deux approches distinctes ont été utilisées, respectivement pour les variables numériques et catégoriques. Pour les premières, la corrélation point-bisériale de Pearson a été calculée, tandis que pour les secondes, le test du χ^2 a été réalisé. Dans le deux cas, c'est l'implémentation du test statistique disponible dans la librairie *Scipy* qui a été utilisée [6].

Les données numériques et textuelles ont d'abord été séparées au moyen de la méthode *infer_ objects()* de la classe *DataFrame* de la librairie *pandas*. Parmi les données textuelles, toutes celles comptant plus de 24 valeurs uniques ont été considérées comme du texte libre et retirées, tandis que les autres ont été conservées en tant qu'attributs catégoriques. Ceux parmi ces derniers ne comptant que deux catégories (incluant d'éventuelles valeurs manquantes) ont été réencodés sous forme de vecteurs *one-hot*.

Les corrélations point-bisériales de Pearson ont d'abord été calculées pour chacun des attributs numériques. Ceci a été fait séparément pour chacune des possibilités d'intention de vote. À chaque fois, le vecteur d'intentions de vote a été réencodé sous forme *one-hot*, où 1 était assigné à la valeur d'intérêt alors que toutes les autres étaient changées en 0.

Le test du χ^2 a ensuite été utilisé afin d'évaluer l'association entre chaque attribut catégorique (incluant les vecteurs *one-hot*) et les différentes possibilités d'intention de vote. Cela a été fait suivant la même approche que pour les attributs numériques, en réencodant le vecteur d'intentions de vote selon la valeur considérée. Pour ces attributs catégoriques, une association globale a aussi été mesurée sans faire de distinction entre les possibilités d'intention de vote. À cette fin, le test du χ^2 a été effectué à partir d'une table de contingence des différents niveaux du vecteur d'intentions de vote et de chacun des attributs catégoriques. Ces variables ont ensuite été classées par ordre décroissant de leur statistique du test du χ^2, afin de guider les décisions futures.

Puisqu'un très grand nombre de comparaisons statistiques ont été effectuées, une correction de Bonferroni a été appliquée aux valeurs p obtenues autant pour les corrélations que pour les tests du χ_2, afin de réduire le nombre d'associations faussement significatives. Cette correction a été réalisée à l'aide de son implémentation tirée de la librairie *statsmodels* [7].

2.4. **Corrélations et distances entre les attributs**

Afin d'évaluer la similarité entre les attributs numériques, nous avons calculé le coefficient de corrélation de Pearson entre chacun d'eux, par paires. L'implémentation *Scipy* a encore une fois été utilisée [6].

Pour les attributs catégoriques, une autre approche devait être utilisée. Nous voulions en particulier évaluer dans quelle mesure les différents vecteurs *one-hot* contenaient la même information. Si deux questions étaient équivalentes, chacun des participants donnerait une même réponse pour chacune, de sorte que les deux vecteurs correspondants contiendraient la même séquence de 0 et de 1. La distance de Hamming, qui mesure le nombre de positions qui diffèrent entre deux séquences de même longueur, semblait donc particulièrement appropriée. Celle-ci a été calculée par paire entre tous les vecteurs *one-hot* n'ayant pas été retirés lors de la préparation du jeu de données. L'implémentation *Scipy* de cette mesure a été utilisée [6].

2.5. **Distances entre les choix de parti**

Nous avons aussi voulu évaluer dans quelle mesure les préférences de parti d'un même répondant pouvaient varier d'un niveau à l'autre (intention de vote, parti auquel il s'identifie, vote de 2015 et parti abordant le mieux l'enjeu le plus important identifié par l'électeur). La distance de Hamming semblait encore une fois très appropriée, mais les quatre attributs correspondants (intention de vote composite, *cps19_fed_id*, (*cps19_vote_2015*), (*cps19_imp_iss_party*)) étaient catégoriques. Chacun a donc d'abord été réencodé de façon à devenir une séquence de caractères: chaque possibilité de parti a été remplacé par un chiffre de 0 à 8. Les distances de Hamming ont ensuite été calculées par paire entre chacun des vecteurs ainsi obtenus [6].

2.6. **Analyse factorielle**

L'analyse factorielle (AF), est un algorithme de réduction de dimensionnalité. Le principe repose sur la recherche de facteurs qui généralisent plusieurs variables. On détermine leur corrélation en utilisant la matrice des corrélations.
Nous avons utilisé la librairie de scikit-learn[8]. Pour mettre en évidence l'éventuelle corrélation qui existerait entre les différents attributs, nous avons d'abord présélectionné des attributs qui seraient plus susceptible de contenir de l'information. En effet, tous les attributs apportant aucune information à nos données (*e.g "Unnamed : 0"*) ainsi que ceux à réponse libre ont été retirés, tel que précédemment. Les attributs trop lacunaires (plus de 1000 données manquantes) ont par la suite aussi été retirés, à l'exception des vecteurs *one-hot*, dont les valeurs manquantes avaient déjà été réencodées en 0. Une telle démarche ajoute nécessairement du biais dans notre analyse, mais ce biais-là est difficilement évitable car beaucoup d'informations sont manquantes. Finalement, une fois les premiers attributs retirés, nous avons retiré les attributs dont l'ensemble des réponses est un *singleton* : ces attributs n'apportent plus d'informations puisqu'ils ne mettent pas en évidence des différences entre les individus qui ont répondu au sondage. Ainsi, sur ces données, l'Analyse Factorielle exploratoire a pu être utilisée pour mettre en évidence les valeurs propres en fonction des différents facteurs (*Figure 4*).

L'objectif dans un premier temps, était de sélectionner le nombre de facteurs qui représenterait le mieux l'ensemble de nos données. Pour ce faire, nous avons utilisé le critère de Kaiser, qui consiste à prendre uniquement les facteurs dont la valeur propre est supérieure à 1. Nous en avons donc conclu que 40 facteurs sont amplement suffisants pour représenter l'intégralité de nos données. Il restait donc à déterminer les corrélations existant entre chaque facteur et les attributs de départ.

3. **Résultats et discussion**

3.1. **Analyses préliminaires des données**

Comme première étape, nous avons calculé le coefficient de variation pour chaque attribut selon le parti pour lequel l'individu a voté. Cette étape a révélé des données aberrantes dans les colonnes *cps*19_*household* et *cps*19_*income_number*. Le coefficient de variation pour ces deux attributs est en effet très élevé, dépassant de plus d'un ordre de magnitude ce qui est observé pour les autres variables. En étudiant les données, nous avons découvert des individus ayant répondu des nombres impossibles, par exemple un revenu de $6.75E+60$\$ en 2018 ou bien une taille de ménage de 7766666 personnes. Nous enlèverons donc ces individus de notre jeu de données pour l'entraînement.

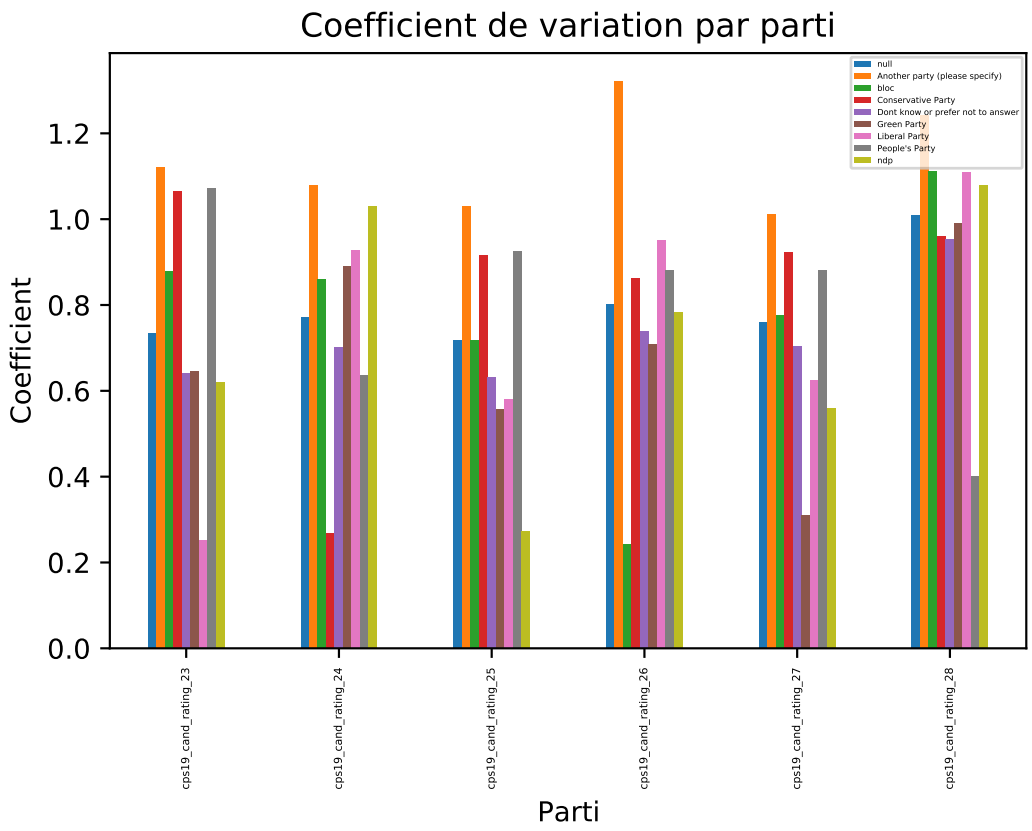

Figure 1. **Les coefficients de variation sont plus bas pour les questions qui quantifient la position de l'électeur par rapport au parti pour lequel il souhaite voter**

Nous avons aussi constaté que le coefficient de variation est plus bas lorsque l'attribut est associé au parti pour lequel la personne vote. Par exemple, l'attribut $cps19\_cand\_rating_23$ a un coefficient de variation nettement plus bas pour les libéraux, car cet attribut représente l'appréciation de l'individu envers le candidat libéral local. Il est donc attendu que la plupart des libéraux donnent une note de 100 à cette question et que les autres donnent des notes plus variables (souvent plus basse). Avec cette information, on peut prédire que les attributs demandant de noter les chefs, les candidats et les partis seront de bons indicateurs du parti pour qui l'individu va voter.

Nous avons ensuite déterminé le mode pour les attributs nominaux et ordinaux. Puisque la plupart des attributs sont des choix de réponses divisés en choix unique (un attribut est donc soit une réponse unique, soit nul), le mode n'est pas très révélateur pour la plupart des questions. Cependant, nous avons pu déceler certaines propriétés intéressantes dans le jeu de données.

Attribut	Another party	Bloc	Conservative	Don't know	Green	Liberal	PPC	NPD
cps19_province	Ontario	Quebec	Ontario	Ontario	Ontario	Ontario	Ontario	Ontario
cps19_imp_iss_party	Another party	Bloc	Conservative	Don't know	Green	Liberal	PPC	NPD
cps19_lead_rating_25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0
cps19_age	37	62	62	58	33	66	32	29

Le mode de l'attribut *cps*19_*province* est "Ontario" pour chaque parti, sauf pour le Bloc Québécois, qui a comme mode "Quebec". Ce n'est pas surprenant étant donné que seul les Québécois peuvent voter pour le Bloc Québécois, mais cela indique que cet attribut pourrait donner une idée initiale à savoir si la personne votera pour le Bloc Québécois ou non. Pour les attributs où l'individu doit choisir le meilleur parti ou donner une note pour chaque parti, il n'est pas surprenant de voir que le mode coïncide avec l'intention de vote de l'électeur. Par exemple, le mode pour les notes données au chef du NPD par les électeurs du NPD est de 100, alors qu'il est de 0 pour les chefs des autres partis. Ceci renforce l'hypothèse précédente qui indiquait que les attributs où l'électeur doit donner une note aux partis seront révélateurs des intentions de vote. On voit aussi une bonne division au niveau de l'âge des électeurs selon les partis. Les libéraux, conservateurs et bloquistes ont un mode d'âge en haut de 62 ans, alors que les autres partis ont des électeurs plus jeunes.

Étant donné que notre objectif est la prédiction des intentions de vote composites (Méthodes), il semblait primordial d'élucider les relations unissant les différents attributs de l'ensemble de données à celles-ci. À cette fin, nous avons calculé les corrélations entre les attributs tirés de l'enquête préélectorale et chacune des possibilités d'intention de vote, suivant deux approches selon la nature des prédicteurs potentiels (*Figure 2*, voir Méthodes). Ces analyses ont d'abord révélé que plusieurs attributs sont fortement corrélés aux intentions de vote des participants. C'est notamment le cas des séries *cps19_party_rating*, *cps19_lead_rating* et *cps19_cand_rating*, qui montrent que les répondants ont tendance à percevoir positivement le parti pour lequel ils souhaitent voter de même que son chef et ses candidats, tout en ayant une opinion négative des partis qui lui sont opposés (*Figure 2 A.*). Cela est particulièrement visible entre les libéraux et les conservateurs, de même qu'entre ces derniers et les néo-démocrates. Similairement, les participants tendent à attribuer des caractéristiques positives bien plus souvent au chef pour lequel ils souhaitent voter qu'à ses adversaires, comme l'indiquent les attributs de la série *cps19_lead* (*Figure 2 B.*). Ceux-ci dénotent notamment si chaque chef est intelligent, digne de confiance, etc. Il s'agit évidemment du genre de patrons auxquels on pouvait s'attendre intuitivement. Ces analyses de corrélation ont aussi mis en évidence plusieurs attributs qui sont significatifs pour la totalité ou presque des intentions de vote possibles. C'est entre autres le cas de *cps19_fed_id*, *cps19_vote_2015*, *cps19_imp_iss_party* et de la série *cps19_issue_handle*. Ceux-ci correspondent au parti auquel les répondants s'identifient, au vote de chaque individu à l'élection de 2015, de même qu'au parti abordant le mieux une gamme d'enjeux électoraux, dont celui que chaque participant considère comme le plus important. En plus de tels attributs d'intérêt général, ces analyses permettent aussi d'identifier des attributs discriminant un seul parti. Par exemple, les partisans du NPD sont en général plus jeunes (*cps19_yob*, *cps19_age*, *Figure 2 A.*) et les indécis (Ne savent pas ou préfèrent ne pas répondre) ont moins d'intérêt envers l'élection et la politique en général (colonnes *cps19_interest*, *Figure 2 A.*), tandis que les partisans du Bloc Québécois sont plus nombreux à se dire québécois et à parler français tout en se distinguant par leur province d'origine (*cps19_ethnicity_29*, *cps19_language_69*, *cps19_province Figure 2 B.*).

Les patrons en escalier visibles dans ces matrices suggèrent l'existence de corrélations entre plusieurs attributs de l'ensemble de données. Nous avons donc étudié celles-ci de manière plus rigoureuse. Nous avons d'abord corrélé l'opinion (positive, neutre ou négative) que les participants ont des partis, de leur chef et de leurs candidats. Sans surprise, les valeurs récoltées pour un même parti présentent une forte corrélation positive, de l'ordre d'environ $0,75$ (*Figure 5 A.*). Tel que suggéré précédemment par l'étude des intentions de vote, on observe aussi une corrélation positive – de moindre magnitude – entre l'opinion d'un parti et la perception de ses chances de victoire aux échelles nationale et locale (séries *cps19_most_seats* et *cps19_win_local*, *Figure 5 A.*). Les attributs des familles *cps19_party_rating*, *cps19_lead_rating*, *cps19_cand_rating*, *cps19_most_seats* et

cps19_win_local présentent donc une forte colinéarité. On constate aussi une forte corrélation positive entre l'intérêt pour l'élection en cours ainsi que l'intérêt général envers la politique, signe que ces deux attributs sont très similaires l'un à l'autre.

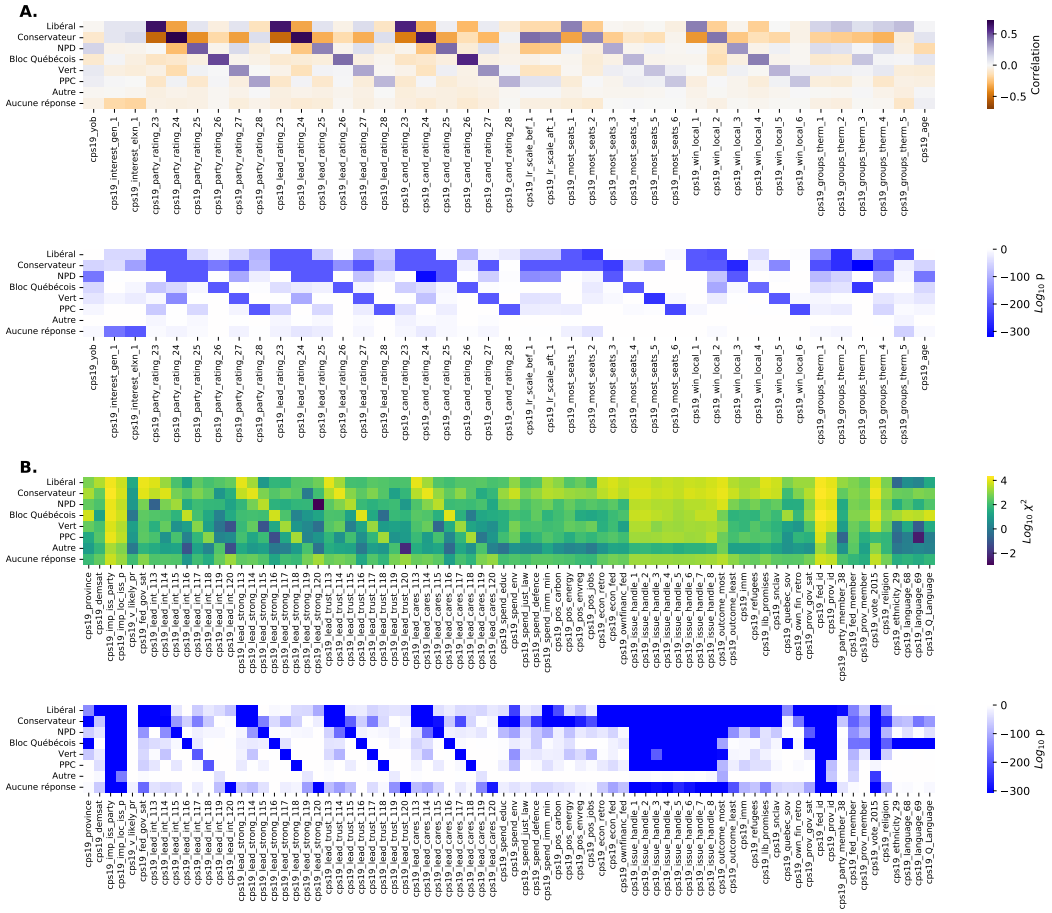

Figure 2. **Plusieurs attributs numériques et catégoriques sont corrélés aux différentes possibilités d'intention de vote. A.** Corrélations point-bisériales entre des attributs numériques sélectionnés et les intentions de vote des participants. Coefficients de corrélation (haut) et Log_{10} de la valeur p correspondante (bas). **B.** Association d'attributs catégoriques sélectionnés (incluant des vecteurs *one-hot*) aux intentions de vote. Log_{10} de la statistique du test du χ^2 (haut) et Log_{10} de la valeur p correspondante (bas). Des attributs en apparence moins informatifs ont été retirés des matrices afin de réduire la taille des figures, mais ne sont pas nécessairement exclus des analyses subséquentes.

Du côté des variables catégoriques, nous avons aussi voulu savoir à quel point les quatre questions de la série *cps19_lead* présentent la même information. Puisqu'il s'agit de vecteurs *one-hot*, nous avons calculé la distance de Hamming entre chacun d'eux afin d'évaluer la proportion des réponses d'un même participant qui diffèrent (*Figure 5 B.*). Comme l'illustre les diagonales apparaissant sur la matrice de distances, les répondants à l'Étude ont tendance à accorder toutes les qualités ou pratiquement aucune d'entre elles aux chefs, selon la perception qu'ils ont de ceux-ci. La plupart des vecteurs faisant référence au même leader (par exemple *cps19_lead_int_113* et *cps19_lead_strong_113*, qui concernent le chef du

Parti Libéral Justin Trudeau) ne diffèrent que de 15% à 25%. Ainsi, ces quatre questions contiennent environ la même information quant aux répondants et à leur intention de vote.

Finalement, nous avons aussi voulu évaluer la correspondance entre les intentions de vote et les attributs témoignant d'une préférence pour un parti: le parti auquel le participant s'identifie le plus (*cps19_fed_id*), le parti qui aborde le mieux l'enjeu le plus important aux yeux de chaque répondant (*cps19_imp_iss_party*) et le parti pour lequel chacun a voté en 2015 (*cps19_vote_2015*). Notre raisonnement était qu'un désaccord entre ces attributs pourrait mettre en évidence l'importance de la dimension temporelle des données, en illustrant par exemple des cas d'électeurs néo-démocrates suivant la tendance et optant pour les libéraux malgré leur préférence. Nous avons donc représenté chaque choix de parti par un caractère différent et calculé les distances de Hamming entre les intentions de vote et les trois attributs mentionnés précédemment. Cette analyse a révélé plusieurs désaccords entre les partis préférés et les intentions de vote. Par exemple, celles-ci ne concordent avec *cps19_imp_iss_party* que dans moins de 50% des cas (*Figure 3, gauche*). On observe néanmoins une très bonne correspondance entre les intentions de vote et le parti auquel les répondants s'identifient le plus, de l'ordre de 70%, et celle-ci atteint même les 80% lorsque les électeurs indécis (ou préférant ne pas répondre) sont retirés des données (*Figure 3, droite*).

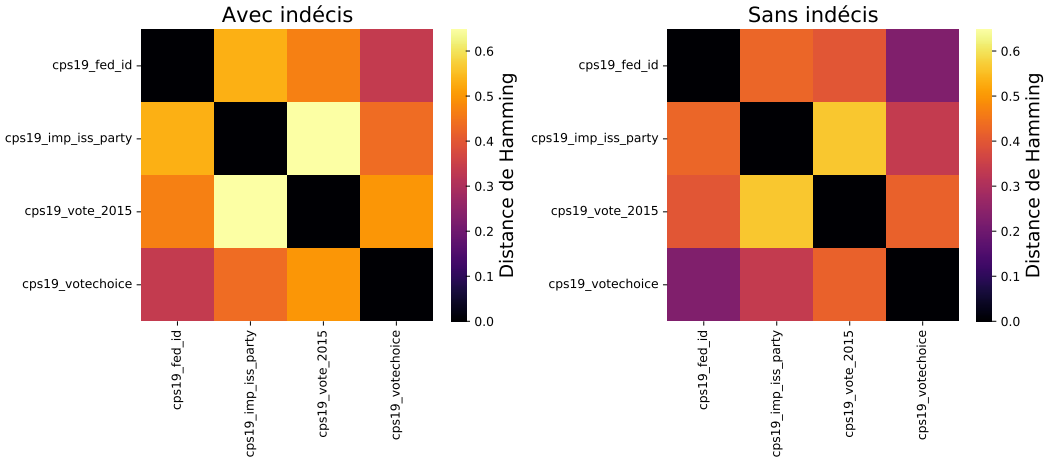

Figure 3. **Les intentions de vote et les partis préférés diffèrent dans la majorité des cas, mais les répondants indécis ont un impact important.** Distance de Hamming par paires entre les intentions de vote et les différentes préférences de parti, selon que les répondant indécis (ne savent pas ou préfèrent ne pas répondre) soient considérés (gauche) ou exclus (droite). L'identifiant *cps19_votechoice* désigne les intentions de vote composites obtenues précédemment (Méthodes), et non l'attribut du même nom de l'ensemble de données de départ.

3.2. **Choix des attributs**

Les analyses précédentes permettent déjà d'identifier des attributs candidats en vue d'une éventuelle prédiction des intentions de vote. Deux séries d'attributs sont tout d'abord particulièrement intéressantes, puisqu'elles créent des patrons caractéristiques en escalier discriminant entre chacun des six principaux partis tout en étant associées aux corrélation de plus grande magnitude et/ou aux valeurs p les plus petites. Du côté des attributs numériques, il s'agit de la famille *rating* incluant les *cps19_party_rating*, *cps19_lead_rating* et *cps19_cand_rating*. Ces trois séries étant fortement corrélées entre elles (*Figure 5 A.*), il semble plus judicieux d'en conserver une seule. Nous privilégierions *cps19_party_rating*, étant donné qu'elle présente, dans l'ensemble, les plus fortes corrélations. Une moyenne des

deux séries pourrait aussi être pertinente: les deux approches mériteraient d'être comparées. Du côté des attributs catégoriques, les quatre séries *cps19_lead* semblent tout aussi prédictives. Celles-ci sont aussi très similaires les unes des autres (*Figure 5 B.*), de sorte qu'il serait préférable d'en conserver une seule. Le classement en ordre décroissant des statistiques du test du χ^2 (non montré dans le présent document) suggère que la série *cps19_lead_cares* soit le plus corrélée aux intentions de vote: c'est donc celle dont nous nous servirions en priorité. On note cependant que les meilleures séries d'attributs numériques et catégoriques que nous venons d'identifier sont elle-mêmes fortement corrélées l'une avec l'autre, puisqu'elles sont associées à des patrons très similaires (*Figure 2 A. et B.*). Afin de limiter la colinéarité entre les attributs utilisés dans le cadre de notre approche d'apprentissage, une seule devrait probablement être conservée. Intuitivement, la série *cps19_party_rating* semble être la plus informative, puisque le fait qu'il s'agisse d'une variable numérique allant de 0 à 10 plutôt que d'une variable dichotomique offre une plus grande résolution.

Il serait judicieux d'ajouter à cette première série d'attributs certaines des variables qui présentent une association forte pour toutes les possibilités d'intention de vote. La plus pertinente est *cps19_fed_id*, dénotant l'identification à l'un ou l'autre des partis, étant donné qu'elle est identique à environ 70% à l'intention de vote composite que nous souhaitons prédire (*Figure 3*). Les attributs *cps19_vote_2015* et *cps19_imp_iss_party* sont aussi très informatifs, mais leur similitude avec *cps19_fed_id* pourrait poser problème. Puisque *cps19_imp_iss_party* ne présente qu'environ 45% d'identité avec *cps19_fed_id*, nous envisagerions de l'inclure si la performance du modèle n'est pas suffisante.

Afin de prédire les intentions de vote le plus fidèlement possible, des attributs spécifiques à certains partis devraient aussi être inclus. On peut penser à *cps19_ethnicity_29* et *cps19_language_69*, qui identifient les Québécois et les francophones, pour le Bloc Québécois, à *cps19_age* pour le NPD, à *cps19_issue_handle_5* pour le PCC, puisque ses électeurs sont plus nombreux à juger qu'il s'agit du meilleur parti pour traiter l'enjeu de la défense nationale, et de l'intérêt envers l'élection (*cps19_interest_elxn_1*) pour les indécis. La province d'origine des participants pourrait aussi être un ajout intéressant, puisqu'elle identifie les conservateurs et les bloquistes. De plus, même si nous avons pour l'instant décidé de l'ignorer, nous reconnaissons que la dimension temporelle (date de soumission des questionnaires) pourrait être informative, puisque les intentions de vote fluctuent en cours de campagne. La très grande correspondance entre l'intention de vote composite et *cps19_fed_id* (*Figure 3*) suggère que la préférence pour un parti capture à elle seule ces variations, mais nous envisageons de revisiter cet aspect si la performance de notre modèle est insuffisante.

Afin de valider ces analyses, et surtout de pouvoir procéder au choix des attributs de façon encore moins arbitraire, nous les avons combinées à une autre approche: l'analyse factorielle (AF).

Celle-ci nous a permis d'obtenir un tableau mettant en évidence la corrélation entre chacun de nos facteurs et les attributs initiaux. Plus la valeur absolue du score des attributs est élevée, plus ces attributs en question sont corrélées au facteur. Chaque facteur correspond quant à lui à un attribut implicite plus général que l'on peut déduire. Cette approche permet donc de trouver un concept reprenant l'idée de chacun des attributs corrélés à nos facteurs.

En l'occurrence dans notre cas, nous avons obtenu une série de corrélations dont la liste suivante est un extrait:

Facteur 1	cps19_lead_int_113	cps19_lead_strong_113	cps19_lead_trust_113	cps19_lead_cares_113	...
Facteur 2	cps19_lead_trust_116	cps19_lead_cares_116	cps19_ethnicity_29	cps19_ethnicity_38	...
Facteur 3	cps19_lead_int_114	cps19_lead_strong_114	cps19_lead_trust_114	cps19_lead_cares_114	...
Facteur 4	cps19_lead_int_115	cps19_lead_strong_115	cps19_lead_trust_115	cps19_lead_cares_115	...
Facteur 5	cps19_lead_int_117	cps19_lead_strong_117	cps19_lead_trust_117	cps19_lead_cares_117	...
Facteur 6	cps19_lead_int_116	cps19_lead_strong_116	cps19_lead_trust_116	cps19_lead_cares_116	...
Facteur 7	cps19_lead_int_118	cps19_lead_strong_118	cps19_lead_trust_118	cps19_lead_cares_118	...
Facteur 8	cps19_party_member_36	cps19_party_member_37	cps19_party_member_39	cps19_fed_donate	...
Facteur 9	cps19_ethnicity_25	cps19_language_72			
Facteur 10	cps19_property_3	cps19_property_4	cps19_property_5		

L'AF permet donc l'analyse des données, afin de mettre en évidence des relations entre les variables dont on ne se serait pas nécessairement doutés. En plus de cet aspect d'analyse, il est possible de réduire la dimension des données afin d'appliquer un algorithme de classification. En effet, l'analyse factorielle permet de réaliser des projections des données sur chaque facteur.

Toutefois, il n'est pas toujours intéressant d'utiliser la réduction de dimensionnalité. En effet, certains algorithmes tels que le SVM, dépendent fortement de la dimensionnalité des données. De ce fait, réduire la dimension des données entraînerait une baisse de précision et de performance dans un tel cas.

Il faut donc veiller à ce que la réduction de dimensionnalité soit compatible avec le type d'algorithme utilisé. Par ailleurs, l'utilisation d'un trop grand nombre de dimensions a pour conséquence de rendre la tâche de classification bien plus laborieuse, et ce, en raison du fléau de dimensionnalité.

En adoptant deux approches distinctes pour la sélection des données, nous avons pu mettre en évidence des corrélations entre les attributs et la pertinence de chaque attribut. L'AF n'est pas en contradiction avec les analyses précédentes de corrélation, mais vient plutôt confirmer et étoffer les patrons découverts, tout en révélant des nuances qui n'avaient pas été décelées par l'approche plus rudimentaire.

Nous avons notamment pu mettre en exergue, dans nos deux approches distinctes, le fait que les *cps19_lead_cares* sont des données relativement importantes et corrélées. Elles font notamment partie des composantes des cinq premiers facteurs dont la valeur propre est assez élevée (signe de l'importance de ces attributs dans nos données). L'AF confirme de plus les corrélations entre les attributs concernant un même chef ou parti, tout en détaillant celles-ci. Par exemple, le facteur 1 regroupe la perception du premier ministre sortant et chef du Parti libéral, de même que des indicateurs de satisfaction envers son gouvernement (*cps19_lib_promises* et *cps19_snclav*) et l'impact de ses politiques sur l'économie (*cps19_econ_fed*). Le facteur 2 vient aussi confirmer l'association décrite précédemment entre l'identité québécoise et une perception positive du Bloc Québécois, puisque l'opinion de son chef, la langue (française ou anglaise), l'ethnicité québécoise ou canadienne-française et la province de résidence s'y trouvent combinés. Cela confirme donc, sans surprise, que l'identité québécoise aide à identifier efficacement les électeurs du Bloc. Finalement, on constate aussi que la perception positive du chef conservateur ainsi qu'une désapprobation globale de la classe politique (opinion positive d'aucun chef) se retrouvent ensemble au sein du facteur 3. Ceci révèle un cynisme et/ou une fatigue politique des électeurs de droite qui nous avait échappée et pourrait aider à identifier les indécis ou les partisans d'autres partis non nommés dans le questionnaire.

L'AF, au delà d'être un outil de réduction de dimensionnalité, se révèle ainsi un excellent moyen d'analyser et d'observer nos données sous un autre angle afin de confirmer nos choix d'attributs.

3.3. **Algorithme d'apprentissage**

Étant donné le gain d'information apporté par certains attributs, par exemple la corrélation entre le fait qu'un électeur soit québécois et ses chances de voter pour le Bloc Québécois,

nous croyons qu'un classificateur de type *DecisionTree* (arbre de décision) suffirait pour bien classifier les données. Ce type d'algorithme est effet en mesure d'apprendre efficacement une telle règle de décision, stipulant par exemple que les répondants votant pour le Bloc résident nécessairement au Québec. De plus, son processus de décision en arborescence pourrait faciliter l'identification des partisans d'autres partis mineurs, puisque ceux-ci présenteront une grande variabilité et se trouveront autant à la gauche qu'à la droite du spectre idéologique. Si les résultats de nos entraînements avec le *DecisionTree* sont insatisfaisants, nous tenterons l'expérience avec un classificateur de type *RandomForest* (forêt aléatoire), qui est un modèle plus complexe mais plus puissant. Il entraîne plusieurs arbres sur des parties différentes du jeu de données et prédit la classe selon la moyenne de tous ces arbres.

Si nous utilisons la librairie scikit-learn[8], nous devrons transformer les attributs catégoriques en vecteurs *one-hot*, car celle-ci ne prend pas en charge les attributs catégoriques. Nous utiliserons des vecteurs ainsi encodés car c'est une façon simple de diviser les attributs catégoriques sans leur donner de valeurs ordinales qui pourraient fausser nos résultats.

Si nous utilisons la librairie H20 [9], nous n'aurons pas à transformer les données, car celles-ci sont prises en charge.

Les arbres de décisions, qui sont aussi à la base des algorithmes *RandomForest*, fonctionnent en divisant le jeu de données actuel en deux, au point où le gain d'information est le plus élevé. Cette étape est effectuée de manière récursive jusqu'à ce que la profondeur de l'arbre maximale soit atteinte ou jusqu'à ce que chaque individu soit bien classé. Étant donné que nous avons seulement gardé des attributs avec un grand gain d'information pour chaque parti, nous croyons qu'un algorithme basé sur les arbres de décision performera bien. Nous devrons donner un maximum à la profondeur de l'arbre, car sinon, il y aura un grand risque de surentraînement, à cause des conditions de fin énumérées ci-haut.

Quel que soit le choix final d'algorithme, la présence de données manquantes peut se révéler problématique. Étant donné que la presque totalité des attributs informatifs que nous avons sélectionnés (Section 3.2) comptent plus de 20 000 réponses, nous pensons que la plupart de celles-ci pourront simplement être ignorées. Il y a néanmoins deux exceptions importantes à cela: *cps19_issue_handle_5* (16 491 réponses) et tous les attributs de la série *cps19_rating* qui concernent le Bloc (répondus seulement au Québec). De plus, toute exclusion des répondants avec données manquantes devra préalablement être validée. Si cela modifie la fréquence des classes et/ou la distribution des prédicteurs à l'intérieur de chacune des classes, les données manquantes devront être remplies pour éviter de créer un biais. Une telle manipulation peut être hasardeuse – risquant elle-même d'introduire de nouveaux biais – et nous privilégions une approche par k plus proches voisins (*KNN*) afin de préserver les caractéristiques de chaque classe. Si la plupart des données devraient être manquantes au hasard (questions non posées à certains participants), certaines ne le seront pas puisque le refus de répondre ou l'indécision y résultent en une donnée manquante (attributs numériques de la série *cps19_rating*). Le remplissage des données manquantes devra donc être abordé avec prudence.

3.4. **Régime d'entraînement**

Pour commencer, nous devrons équilibrer notre jeu de données, car la proportion de classes est très en faveur des libéraux et des conservateurs (voir plus bas).

Pour cet équilibrage, nous devrions faire un mélange de sur-échantillonnage pour les classes sous-représentées et du sous-échantillonnage pour les classes sur-représentées. Puisque nous n'utiliserons pas beaucoup d'attributs pour entraîner notre classificateur, il est dangereux de faire du sur-échantillonnage et d'avoir des données insensées ou des données qui se répètent. Étant donné que la classe la moins représentée a une proportion de moins de

1% du jeu de données, le sous-échantillonnage enlèverait trop de bonnes données dans les classes mieux représentées.

Parti	Proportion (%)
Liberal	28.4604
Conservative	26.9454
Don't know/Prefer not to answer	15.6086
NPD	13.8416
Green	7.9273
Bloc	4.5450
PPC	2.0129
Another party	0.6588

Afin de contrer les problèmes énumérés ci-haut, nous pourrions utiliser les deux techniques afin de trouver un juste milieu, soit sur-échantillonner la moitié des données mal représentées et sous-échantillonner le reste.

En assumant que cette méthode nous amènera à avoir un grand nombre de données restantes pour l'entraînement, nous allons entraîner sur 90% des données. Nous utiliserons le 10% restant pour tester le modèle à la fin lorsque l'entraînement donnera de bons résultats.

Pour l'entraînement de l'arbre de décision, nous utiliserons une méthode de validation croisée k-fold à 10 plis. Cette méthode est intéressante car elle rend l'entraînement du modèle plus robuste contre le surentraînement. En entraînant sur le jeu de données au complet, mais en s'assurant d'avoir toujours une partie pour valider, l'algorithme est moins à risque de surentraîner sur une classe précise, par exemple si certaines classes sont moins bien représentées dans un ensemble entraînement/validation mal divisé.

Pour l'entraînement du *RandomForest*, nous utiliserons une division 70/30 pour les données d'entraînement et de validation. Nous n'avons pas besoin d'une approche k-fold, car le *RandomForest* est entraîné avec l'aggrégation *bootstrap*, qui entraîne chaque arbre sur une partie différente du jeu de données.

References

[1] L. Ezrow and G. Xezonakis. "Citizen Satisfaction With Democracy and Parties' Policy Offerings". In: *Comparative Political Studies* 44.9 (2011), pp. 1152–1178. DOI: 10.1177/0010414011405461.

[2] A. Gallego, G. Rico, and E. Anduiza. "Disproportionality and voter turnout in new and old democracies". In: *Electoral Studies* 31.1 (2012). Special Symposium: Germany's Federal Election September 2009, pp. 159–169. DOI: https://doi.org/10.1016/j.electstud.2011.10.004.

[3] W. Jennings and C. Wlezien. "Election polling errors across time and space". en. In: *Nature Human Behaviour* 2.4 (Apr. 2018), pp. 276–283. DOI: 10.1038/s41562-018-0315-6.

[4] P. Hummel and D. Rothschild. "Fundamental models for forecasting elections at the state level". In: *Electoral Studies* 35 (2014), pp. 123–139. DOI: https://doi.org/10.1016/j.electstud.2014.05.002.

[5] L. B. Stephenson, A. Harell, D. Rubenson, and P. J. Loewen. *2019 Canadian Election Study - Online Survey*. 2020. DOI: https://doi.org/10.7910/DVN/DUS88V.

[6] P. Virtanen et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". en. In: *Nature Methods* 17.3 (Mar. 2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[7] S. Seabold and J. Perktold. "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010.

[8] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[9] P. Stetsenko. *Machine Learning with Python and H2O*. Mar. 2020. URL: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/booklets/PythonBooklet.pdf.

Appendix A. Analyse factorielle

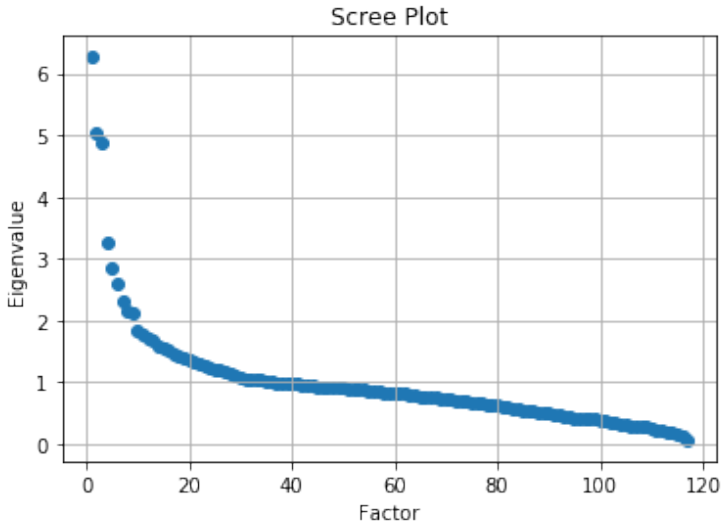

Figure 4. **Valeur propre selon le facteur**

Appendix B. Corrélations et distances entre attributs

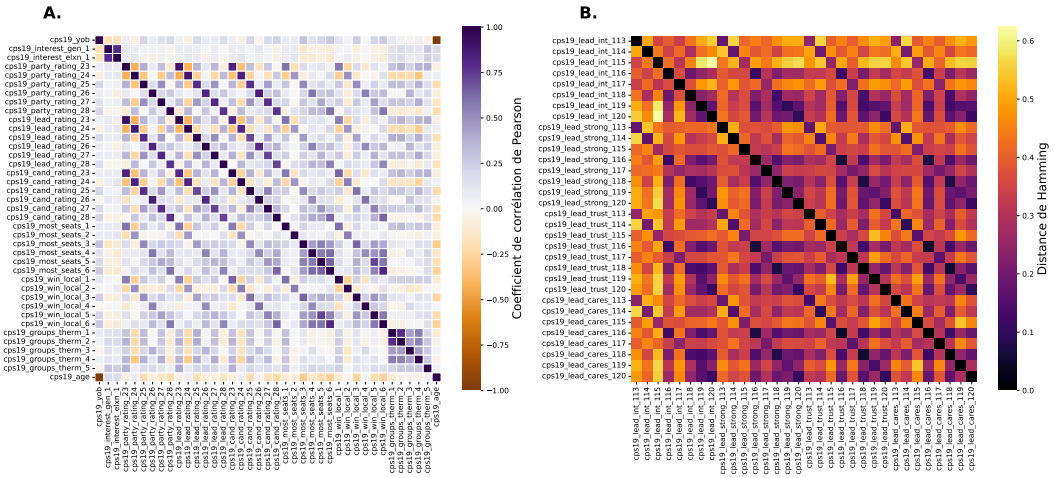

Figure 5. **Certains attributs sont fortement corrélés entre eux. A.** Corrélations de Pearson par paires pour une sélection d'attributs numériques. **B.** Distances de Hamming par paires entre des vecteurs de type *one-hot* sélectionnés.