Approches prédictives pour l'étude de l'effet de pairs : application à la prédiction de préférence pour une spécialité des étudiants en médecine

Mustapha Atmani*, Nathalie Pernelle*, Céline Rouveirol*,
Noemi Berlin**, Magali Dumontet**,
Mathieu Lambotte***
Carole Treibich****

*LIPN CNRS UMR 7030, Université Sorbonne Paris Nord
prenom.nom@lipn.univ-paris13.fr
**EconomiX CNRS UMR 7235, Université Paris Nanterre
nom.prenom@parisnanterre.fr
*** CREM, Université de Rennes
mathieu.lambotte@univ-rennes.fr
****GAEL, Univ. Grenoble Alpes, CNRS, INRAE, Grenoble INP
carole.treibich@univ-grenoble-alpes.fr

Résumé. De nombreux travaux en économétrie ont été développés pour tenter de mettre en évidence un effet de pairs dans une population. Dans de nombreux domaines, des études montrent que les effets des pairs ont un impact important (e.g. science de l'éducation, santé publique, ...). Le modèle le plus utilisé dans ces études empiriques est le modèle linéaire en moyennes, qui suppose que les individus sont affectés de manière linéaire par l'action et les caractéristiques moyennes de leurs pairs. Dans ce papier, nous proposons de comparer les résultats obtenus par deux types d'approches prédictives : (1) un modèle linéaire en moyenne et (2) une approche généraliste de fouille de règles de la logique du premier ordre qui a été adaptée par l'utilisation de biais de langages spécifiques. Les expérimentations ont été menées sur un jeu de données réelles qui concerne le choix d'une spécialité médicale par les étudiants en médecine lorsqu'il doivent sélectionner leur un poste d'interne en fin de leur 6ème année d'étude.

1 Introduction

Un *effet de pairs* peut être défini comme un effet résultant des interactions entre individus d'un même groupe : même environnement (e.g. scolarisés dans la même classe, ou appartenant à un même réseau social), qui peuvent s'influencer à travers leurs caractéristiques et leur comportement. De nombreux travaux récents en économétrie s'intéressent à la mise en évidence d'effets de pairs et produisent des modèles théoriques qui permettent de modéliser le comportement d'un individu comme une fonction à la fois des caractéristiques de cet individu, mais également des caractéristiques du ou des groupes auxquels il appartient (Bramoullé et al.,

2020). De nombreuses études mettent en évidence de fortes corrélations entre la valeur d'une variable d'étude pour un individu (e.g. la réussite scolaire, le fait de pratiquer un sport) et ceux de ses pairs. Découvrir des variables causales est bien plus complexe, car l'existence de corrélations ne démontre bien sûr pas de liens de causalité. En effet, une corrélation observée entre le choix de l'individu et celui de ses pairs peut être causée par des caractéristiques non observées. De plus, il est difficile de distinguer l'impact des comportements des pairs (effets endogènes des pairs) de celui des caractéristiques des pairs (effets contextuels des pairs) en raison de la simultanéité du comportement des agents qui interagissent. Il s'agit du problème dit de 'réflexion' défini par Manski (1993).

Parallèlement, de nombreuses approches de fouille de règles permettent de générer un ensemble des règles représentées en logique du premier ordre (Lajus et al., 2020; Ortona et al., 2018) et de telles représentations peuvent prendre en compte des prédicats décrivant le comportement et les caractéristiques des pairs pour conclure sur la valeur d'une variable d'étude. Cependant certaines règles peuvent ne pas être pertinentes pour ce type d'étude.

Dans cet article, nous nous intéressons au problème suivant : étant donné les données observationnelles d'un réseau, les données relatives aux caractéristiques des individus, les valeurs d'une variable d'étude associée aux nœuds du réseau, l'objectif est de comparer les prédictions obtenues par des approches interprétables : une approche de fouille de règles généraliste et le modèle linéaire en moyenne étendu proposé par Bramoullé et al. (2009). Afin d'adapter l'approche de fouille de règles, nous avons défini différents biais permettant de caractériser les règles pertinentes pour ce type d'étude. Les évaluations réalisées sur un jeu de données décrivant les préférences d'étudiants en 6ème année de médecine montrent que les deux types d'approches réalisent des prédictions de qualité relativement comparables même si les règles générées permettent de prendre en compte des sous-populations définies par des variables explicatives peu significatives dans le modèle statistique. Enfin, les deux approches semblent montrer que les effets de pairs ne sont pas significatifs pour le problème étudié.

2 Etat de l'art

Approches statistiques pour l'étude de l'effet de pairs : Différentes approches ont été développées en économétrie pour tenter d'évaluer les effets de pairs (Bramoullé et al., 2020). A l'origine, les approches se sont focalisées sur l'étude de la valeur d'une variable d'étude dans différents groupes d'individus qui partitionnent une population en fonction d'un critère tel que la ville d'habitation, ou la classe (Manski, 1993). Cependant, les individus ont généralement des ensembles distincts de pairs, avec lesquels ils entretiennent des liens potentiellement non-symétriques, ce qui conduit à considérer plus précisément un réseau d'interactions (e.g. réseaux d'amis ou réseaux sociaux). Aussi, certaines méthodes plus récentes permettent de prendre en compte les réseaux de chaque individu en se basant sur des modèles linéaires en moyenne. (Bramoullé et al., 2009) permet ainsi de modéliser un effet de pairs en distinguant pour une variable d'étude, l'impact des caractéristiques propres de l'individu, de l'influence du choix et des caractéristiques propres des pairs. D'autres approches peuvent également s'appuyer sur des réseaux formés aléatoirement, ou sur des données de panel permettant de différencier les caractéristiques constantes des caractéristiques évolutives (Hanushek et al., 2003).

Dans le cadre de cet article, nous nous intéressons à des observations non évolutives issues de réseaux d'interactions observés (i.e. qui n'ont pas été formés aléatoirement) et pour lesquels des chocs aléatoires ne sont pas appliqués.

Approches de découverte de règles : En apprentissage automatique symbolique et en fouille de données relationnelles, différentes approches de découverte de règles approximatives (i.e. règles IF-Then) de la logique du premier ordre ont été développées qui peuvent permettre de prédire une valeur de variable d'étude (i.e. une classe) pour un individu à partir de données graphes. Les langages utilisés dans certaines de ces approches (restriction de la logique des prédicats) permettent de représenter explicitement des interactions entre individus et donc de modéliser des effets de pairs. Des approches d'apprentissage supervisé en Programmation Logique Inductive se sont intéressées dès les années 1990 à apprendre des règles dans des restrictions de la logique des prédicats. Une attention toute paticulière a été apportée dans les systèmes "historiques" à l'étude des biais de langage pour restreindre l'espace de recherche. Le but était de rendre les systèmes d'apprentissage supervisé génériques adaptables à des données spécifiques (Cropper et Dumancic, 2022; Blockeel et Raedt, 1998). De nombreuses approches plus récentes, issues de la fouille de règles dans les graphes de données, sont généralistes et plus ou moins expressives (e.g. règles limitées aux chemins dans les graphe (Meilicke et al., 2019), contraintes sur les valeurs numériques des prédicats ou règles négatives (Ortona et al., 2018; Nassiri et al., 2023) mais ils proposent peu de biais de langage paramétrables pour un domaine spécifique. AMIE3 (Lajus et al., 2020) est un système de découverte de règles *top-down* qui permet de générer toutes les règles connectées et fermées correspondant à différents biais de langage paramétrables qui sont relatifs à la présence de constantes, au nombre d'atomes en prémisse ou encore aux prédicats qu'il est possible d'utiliser.

A notre connaissance, il n'existe pas d'approche de fouille de règles dédiée à l'effet de pairs. Dans cet article, nous nous intéressons aux approches prédictives prenant en compte les liens du réseau d'amitiés, en espérant que ce type d'approche pourra être ré-utilisée pour rechercher des explications ou des règles hypothétiquement causales (Huang et al., 2021; Simonne et al., 2021).

3 Approche basée sur un modèle linéaire en moyenne

L'objectif est de savoir si le comportement et les caractéristiques des amis ou des personnes avec qui un individu interagit via un lien social influencent le comportement d'un individu. L'un des modèles les plus utilisés en économétrie est le modèle linéaire en moyennes étendu, dans lequel chaque individu est influencé par les caractéristiques moyennes et les valeurs des variables d'études moyennes des personnes de son groupe, qui est défini par ses interactions sociales (Bramoullé et al., 2009). Dans ce modèle, la ou les valeurs des variables d'études sont définies par :

$$y = \alpha + \beta Gy + \gamma x + \delta Gx + \epsilon, \qquad E(\epsilon|x, G) = 0$$

où y est un vecteur de valeurs de variables d'études pour un individu, G est une matrice carrée de dimension $n \times n$, où n est le nombre d'individus dans le réseau. Les éléments de G sont définis comme $G_{ij} = 1/n_i$ si j est un ami de i, et 0 sinon, où n_i est le nombre total d'amis de i, x représente les caractéristiques individuelles (e.g. genre, appétence pour la recherche) et ϵ est un terme d'erreur.

Nous nous focalisons sur le problème de prédiction d'une variable binaire, où la variable d'étude Y prend la valeur 1 si l'individu appartient à une classe d'intérêt, et 0 sinon. Nous avons utilisé un modèle *logit* qui permet de modéliser la probabilité qu'un individu appartienne à une classe en fonction d'un ensemble de variables explicatives.
Significativité de l'impact des variables explicatives : En utilisant ce modèle, nous pouvons estimer l'impact des différentes variables explicatives (i.e. caractéristiques individuelles des individus et des influences de leurs pairs sur la variable d'étude), tout en garantissant que les prédictions restent interprétables en termes de probabilités. La *p-value* associée à chaque coefficient permet de déterminer la significativité statistique des variables explicatives. Si la *p-value* d'un coefficient est inférieure à 0,05, nous pouvons rejeter l'hypothèse nulle et conclure que la variable explicative a un effet statistiquement significatif sur la variable d'étude.
Impact global des variables explicatives issues des pairs : Pour évaluer l'impact de l'effet de pairs sur les capacités prédictives du modèle, nous proposons de comparer le modèle linéaire en moyenne étendu présenté dans cette section avec un modèle, appelé CP, qui prédit l'appartenance à une classe en utilisant uniquement les caractéristiques propres de l'individu.

4 Approche basée sur une approche de fouille de règles

4.1 Présentation de AMIE3

AMIE3 (Lajus et al., 2020) est un système de découverte de règles qui permet de générer un ensemble des règles connectées et fermées à partir de données graphe. Nous considérons ici des graphes de données sociales qui peuvent être définis comme suit :

Définition 1 Graphe RDF de données sociales. *Nous considérons un graphe RDF de données sociales $\mathcal{G}$ qui peut être représenté par un n-uplet $(\mathcal{I}, Ind, P, P_{Res}, L, \mathcal{F})$ où $\mathcal{I}$ désigne l'ensemble des entités, $Ind \subset I$ représente l'ensemble des individus, $\mathcal{P}$ représente l'ensemble des prédicats, $P_{res} \subset P$ désigne l'ensemble des prédicats représentant des liens d'influences potentiels entre individus, et où $\mathcal{L}$ désigne l'ensemble des littéraux (tels que les nombres et les chaînes de caractères). L'ensemble des faits $\mathcal{F}$ est représenté par des triplets qui peuvent s'écrire sous la forme $predicat(sujet, objet)$ tel que sujet $\in \mathcal{I}$, prédicat $\in \mathcal{P}$, objet $\in \mathcal{I} \cup \mathcal{L}$.*

Définition 2 Règle de Horn connectée et fermée. *Une règle de Horn $r : \mathcal{B} \rightarrow H$ est une formule de la logique du premier ordre telle que le corps de la règle $\mathcal{B}$ est une conjonction d'atomes $B_1, ..., B_n$ (i.e. de prédicats $B_i(X, Y)$ où X et Y sont des variables ou des constantes et où la conclusion H est formée d'un atome unique. Une règle est fermée (closed) si chacune des variables apparaît au moins deux fois dans la règle. Une règle est connectée si tous les atomes sont transitivement connectés, deux atomes étant connectés si ils partagent au moins une variable.*

La règle suivante est un exemple simple de règle fermée et connectée : $Classement(?I_1, q1)$, $parents\_medecins(?I_1, vrai) \rightarrow Best\_Spe1(?I_1, autre)$ où les variables sont précédées du caractère ? et où dans cet exemple $q1$, $vrai$ et $autre$ sont des littéraux ($q1, vrai, autre \in L$).

Différentes mesures ont été définies afin d'estimer la qualité de ces règles, mesures qui incluent (ou non) l'hypothèse du monde ouvert (OWA). La couverture hc d'une règle r peut

être définie comme la proportion des instantiations de la conclusion qui sont correctement prédites par la règle. La confiance PCA pca_{conf} de la règle r mesure la précision de la règle dans le cadre de l'hypothèse du monde ouvert (OWA), c'est-à-dire le rapport entre les prédictions correctes et le nombre total de prédictions correctes et incorrectes faites par la règle. En OWA, les prédictions incorrectes sont déterminées en fonction du caractère fonctionnel de la relation apparaissant en conclusion (i.e. si la relation exprimée dans H est plus fonctionnelle qu'inverse fonctionnelle, toute valeur inconnue du 2ème argument appartient aux contre-exemples si une valeur est déjà connue, comme dans l'exemple d'une nouvelle date de naissance).

Nous notons R l'ensemble de règles connectées et fermées qui concluent sur la variable d'étude et telles que la couverture hc est supérieure à un seuil s_{hc} et la valeur de pca_conf est supérieure à un seuil s_{pca}.

4.2 Définition de biais adaptés aux règles pouvant caractériser la présence d'un effet de pairs

En PLI, un biais de langage définit le langage des hypothèses "licites" de l'espace de recherche de l'algorithme d'apprentissage, c'est-à-dire qui sont susceptibles d'être explorées et évaluées sur les données lors du parcours de cet espace. On peut par exemple restreindre les prédicats qui apparaissent dans le corps des règles, le nombre de littéraux utilisant ces prédicats, les types de liens entre les variables de la règle (longueur du plus court chemin de prédicats reliant une variable à une variable apparaissant dans le tête. variables), le nombre de variables existentielles, etc.

AMIE3 permet d'exprimer certains biais de langage et en particulier des biais concernant : (1) l'utilisation de constantes pour un sous-ensemble de prédicats spécifiés, (2) la limitation de la recherche à des règles concluant sur un prédicat donné, ou encore (3) la spécification du nombre maximum d'atomes dans le corps de la règle.

Nous définissons ici deux biais spécifiques à notre étude de l'effet de pairs, qui visent à exclure des règles générées par AMIE3. Tout d'abord, lorsque l'on ne s'intéresse pas à l'étude de l'influence d'individus particuliers, les règles comportant des prédicats traduisant les liens sociaux ne doivent pas être instanciés (e.g. $BestSpe1(id12, ?C), Ami\_de(id12, ?I1) \rightarrow BestSpe1(?I1, ?C)$, où $id12 \in Ind$, qui indique que tous les amis de l'individu $id12$ font le même choix de spécialité que lui n'est pas pertinent). Ce premier biais B_1 est défini sur un ensemble R de règles de la forme $r : B \rightarrow H$:
$B_1(R) = \{r \in R \mid \forall\, B_i(X, Y) \in B$ tel que $B_i \in P_{res}, X \notin Ind$ et $Y \notin Ind\}$

L'objectif du deuxième biais est d'exclure des règles qui ne respectent pas des *chemins d'influence*, en s'appuyant sur des chemins de relations sociales, éventuellement non symétriques de P_{res}. Considérons une relation non symétrique Ami_de et la définition de la relation transitive $Infl : Ami\_de(?a, ?g) \rightarrow Influence(?g, ?a)$.
La règle : $Factspe\_actes\_medtech(?a, 4), Factspe\_precaires(?g, 1), Ami\_de(?a, ?g) \rightarrow Best\_spe1(?a, autre)$ ne doit pas être considérée, car l'influence associée par $Ami\_de(?a, ?g)$ est orientée de l'individu $?g$ vers $?a$, or la prédiction de comportement est faite pour un individu instanciant la variable $?a$ (c'est-à-dire l'influenceur plutôt que l'influencé).

De même, deux individus ne peuvent être liés dans la règle par des valeurs décrivant leurs caractéristiques s'ils ne partagent pas de lien dans le réseau. Ainsi, la règle :
$genre(?I1, ?G), CSP(?I1, 2), genre(?I2, ?G) \rightarrow classement(?I2, b)$ n'est pas pertinente

car les individus $?I1$ et $?I2$ ne sont pas liés par un lien social. Plus généralement, des variables décrivant des individus doivent appartenir à un chemin qui respecte l'influence, définie par $Infl$. Le biais B_2 précisant la contrainte ci-dessus sur un ensemble R de règles de la forme $r : B \rightarrow H$ se définit comme suit :
$B_2(R) = \{r \in R \text{ t.q. } H = p(?Y_0, \_) \mid \forall Y_i \in VarsInd(B), B, Infl \models Influence(Y_i, Y_0)\}$
où $Infl$ est l'ensemble des règles définissan $Influence$ et $VarsInd(B)$ note les variables apparaissant en argument des atomes de B. Nous notons R_{pairs} l'ensemble des règles fermées et connectées qui respectent les biais B_1 et B_2.

L'objectif de cette étude n'est pas de se focaliser uniquement sur les règles comportant au moins un prédicat décrivant un lien social en prémisse. En cohérence avec le modèle linéaire en moyenne étudié, nous proposons d'étudier l'impact de l'effet de pairs sur une approche prédictive à base de règles en comparant les prédictions effectuées avec et sans les règles incluant un lien social. Une règle de R_{pairs} qui comporte au moins un lien social en prémisse est appelé *règle sociale*, et une règle qui ne comporte aucun lien social sera nommée *règle simple* dans les expérimentations.

5 Expérimentations

L'objectif des expérimentations est d'étudier comment l'effet de pairs impacte les prédictions réalisées par les deux types d'approches et de montrer leurs avantages et leurs limites.

5.1 Jeu de données MEDSPE

Depuis plusieurs années, les pouvoirs publics s'inquiètent de l'insuffisance de l'offre de médecins et de leur mauvaise répartition entre les spécialités médicales et sur le territoire. Un des objectifs du projet MEDSPE[1] est de mieux comprendre les déterminants des préférences puis du choix de poste d'interne des étudiants de sixième année de médecine lors de l'attribution des postes. Le jeu de donnée utilisé provient d'une enquête réalisée en mars 2023 auprès d'étudiants en sixième année de médecine (9 278 étudiants) avec plusieurs blocs de questions : préférences des étudiants pour les postes et le classement ECN, importance accordée à 10 caractéristiques dans leur future pratique d'une spécialité, variables comportementales déclaratives (compétitivité, attitudes envers le risque, impatience, altruisme, ...), informations socio-démographiques (age, sexe, le fait d'avoir un médecin dans sa famille, université actuelle), et informations sur l'entourage et le réseau des étudiants ont été recueillies (référence des étudiants avec qui l'étudiant questionné a des affinités, ou avec qui il étudie). Dans le cadre de notre expérimentation, nous cherchons à comprendre comment les préférences et les caractéristiques des amis influencent la décision des étudiants en médecine de choisir la spécialité de médecine générale, specialité pour laquelle il y a le plus grand nombre de postes, par rapport à d'autres spécialités. Il s'agit donc de prédire une variable binaire (*généraliste* ou non).

Initialement, 3011 réponses ont été recueillies. Cependant, nos avons du exclure les réponses qui contiennent des valeurs manquantes. Nous avons également exclu les amis mentionnés qui n'ont pas répondu au questionnaire au moins pour les réponses aux questions correspondant aux variables significatives. L'échantillon final comprend 493 étudiants provenant de 24 universités différentes. Parmi eux, 70 % sont des femmes et 30 % sont des hommes.

1. ANR JCJC N°-21-CE26-0013-01.

En collaboration avec les économistes impliqués dans MEDSPE, 10 variables décrivant les caractéristiques individuelles des étudiants ont été sélectionnées (i.e. suppression des variables non pertinentes et de certaines variables corrélées). Les variables sont les suivantes :

- **Best_spe1** : variable d'étude qui vaut 1 si la préférence de l'étudiant est médecine générale (129 étudiants), 0 sinon (364 étudiants).
- **Sexe_ecn** : le genre de l'étudiant (1 pour les femmes, 0 pour les hommes).
- **Classement_ecn_attendu** : le classement attendu de l'étudiant aux Épreuves Classantes Nationales (ECN) (valeurs variant entre 1 et 9278, recodées en quartiles).
- **Reussite_1ere_ann_med** : Cette variable binaire vaut 1 si l'étudiant a réussi sa première année de médecine du premier coup, 0 sinon.
- **Med_parents** : Cette variable binaire qui prend la valeur 1 si l'un des parents de l'étudiant travaille dans le domaine médical, et 0 sinon.
- **Les facteurs qui caractérisent les spécialités** : Les 6 variables Factspe représentent les dimensions caractérisant l'exercice d'une spécialité qui intéresseraient un étudiant. Les réponses aux questions prennent des valeurs de 1 (pas du tout important) à 5 (extrêmement important) : Factspe_soins_spe (très spécialisée et concentrée sur des traitements/soins spécifiques), Factspe_patients (permet d'avoir beaucoup de contacts avec les patients), Factspe_precaires (soigner les personnes précaires), Factspe_gardes (peu de gardes et astreintes), Factspe_medtech (avec beaucoup d'actes médico-techniques), Factspe_recherche (permet d'avoir des activités de recherche).

5.2 Evaluation du modèle linéaire en moyenne

Dans cette section, nous présentons les résultats obtenus à partir de l'application du modèle linéaire en moyenne pour expliquer la probabilité de choisir la spécialité *généraliste*. Trois modèles distincts ont été testés. Le premier, appelé **Modèle CP**, est un modèle de base qui explique la probabilité de choisir *généraliste* uniquement en fonction des caractéristiques propres de l'étudiant. Le second, le **Modèle de Bramoullé**, introduit en section précédente, prend en compte l'effet des amis dans la prise de décision, en intégrant les caractéristiques propres de l'étudiant ainsi que les caractéristiques moyennes de ses amis. Enfin, le **Modèle de Manski** se distingue du modèle de Bramoullé en considérant la moyenne du groupe d'appartenance, ici l'université, ce qui permet d'analyser l'effet des pairs à un niveau plus global. Ces trois modèles sont comparés afin d'évaluer la pertinence de l'effet de pairs et d'analyser les performances respectives des différentes méthodes.

Estimation des modèles logistiques :

Après avoir défini les variables explicatives utilisées dans chaque modèle, nous avons procédé à l'estimation des paramètres de chaque variable explicative en utilisant un échantillon de 493 étudiants répondants. Cette estimation a été effectuée à l'aide de la fonction `glm()` de R : $model \leftarrow glm(formula, data = medspe, family = binomial(link = "logit"))$
où la variable `formula` est définie comme suit : $Best\_spe1 \sim X_1 + X_2 + ... + X_n$ (X_1, X_2, ..., X_n représentant les différentes variables explicatives utilisées dans chaque modèle).

Dans ce qui suit, nous présentons un extrait des résultats obtenus à partir de l'estimation des coefficients de chaque modèle. Les tableaux 1 et 2 incluent uniquement les variables significatives, i.e. celles dont la *p-value* est inférieure à 0.05. Le signe négatif du coefficient estimé pour la variable sexe_ecn indique par exemple que les hommes ont plutôt tendance à choisir

généraliste tandis que le signe positif pour factspe_patient indique que la volonté d'avoir des contacts avec les patients est en faveur de *généraliste*. Les coefficients estimés sont accompagnés de leur odds ratio qui permet de quantifier l'effet des variables et de comparer leur impact relatif sur la probabilité de choisir la spécialité *généraliste*. Un odds ratio supérieur à 1 indique que l'augmentation de la variable explicative est associée à une probabilité accrue de choisir *généraliste*, tandis qu'un odds ratio inférieur à 1 suggère une relation inverse.

TAB. 1 – *Coefficients estimés et Odds Ratios pour le modèle CP pour les variables significatives du modèle*

Variable	Estimation	Odds Ratio	P-value
Sexe_ecn	-0.686	0.504	0.0256 *
Med_parents	-1.151	0.317	0.0310 *
Factspe_soins_spe	-0.768	0.464	1.22e-05 ***
Factspe_patients	0.526	1.693	0.0041 **
Factspe_gardes	0.732	2.080	1.82e-08 ***
Factspe_recherche	-0.506	0.603	0.0012 **

Le tableau des résultats du modèle CP met en évidence plusieurs variables propres significatives, telles que $Sexe\_ecn$ et $Factspe\_gardes$, qui influencent la probabilité de choisir une spécialité *généraliste*. Dans le modèle de Bramoullé, bien que plusieurs variables soient significatives, aucune d'entre elles ne reflète directement l'effet des pairs, ce qui ne permet pas de montrer une influence claire des choix et des caractéristiques moyennes du cercle d'amis au sein de l'université.

Les variables significatives pour le modèle de Manski sont omises dans l'article, ce sont les mêmes que celles des autres modèles avec deux variables supplémentaires :
$Classement\_ecn\_attendu$ (cette variable code le quartile du classement attendu de l'étudiant aux ECN, voir section 5.1) avec un odds ratio de 1.521 et une *p-value* de 0.004934 et surtout la variable $mBest\_spe1$ (proportion des étudiants de l'université ayant choisi la spécialité *généraliste*) (odds ratio de 5.2678e+03 et *p-value* de 3.60e-06) est particulièrement significative, suggérant que la préférence pour cette spécialité médicale pourrait être fortement influencée par les choix collectifs au niveau de l'université.

TAB. 2 – *Coefficients estimés et Odds Ratios pour les variables significatives du modèle de Bramoullé*

Variable	Estimation	Odds Ratio	p-value
Sexe_ecn	-0.930	0.372	0.00751 **
Factspe_soins_spe	-0.752	0.476	3.39e-05 ***
Factspe_patients	0.548	1.722	0.00405 **
Factspe_precaires	0.381	1.464	0.00772 **
Factspe_gardes	0.779	2.180	5.09e-09 ***
Factspe_recherche	-0.520	0.595	0.00101 **

Evaluation et comparaison des modèles : Nous avons évalué les 3 modèles en utilisant une 10 validation croisée stratifiée. Les figures suivantes présentent les courbes ROC moyennes avec écart-type pour les trois modèles. Chaque graphique montre la capacité du modèle à différencier les classes positive et négative pour des seuils de classification variant de 0 à 1, ainsi que la variabilité des performances mesurée par l'écart-type.

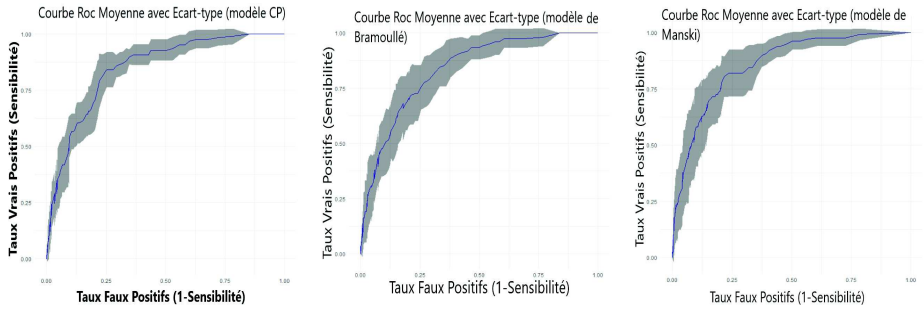

FIG. 1 – *Courbe ROC moyenne pour le modèle CP (à gauche), le modèle de Bramoullé (au centre), le modèle de Manski (à droite)*

On observe que la surface grise (écart-type) varie légèrement d'un modèle à l'autre. Les modèles de Bramoullé et Manski présentent un plus grand écart-type sur les plis, tandis que le modèle CP a un écart-type plus faible, suggérant une performance plus stable au travers des différents seuils de classification. L'aire sous la courbe ROC correspondant au modèle CP est légèrement plus importante, ce qui traduit un meilleur comportement du modèle en général.

Afin d'évaluer les performances des modèles en terme de rappel, précision et F-mesure, nous avons identifié le meilleur seuil de classification $s \in [0, 1]$ pour chaque modèle, (i.e si $y \geq s$, $Best\_spe1 = generaliste$) et avons sélectionné celui qui maximise la justesse moyenne sur l'ensemble des plis : 0.4 pour le modèle CP, 0.5 pour le modèle de Bramoullé, et 0.5 pour le modèle de Manski. Les résultats sont présentés dans le tableau 3 ci-dessous, présentant pour chaque modèle la précision moyenne, le rappel moyen, la F-mesure moyenne, ainsi que la justesse accompagnés de leurs écart-types respectifs.

TAB. 3 – *Résultats de la 10-validation croisée pour chaque modèle statistique : moyenne (± écart-type) sur les 10 plis*

Modèle	**Précision**	**Rappel**	**F-mesure**	**Justesse**
CP	0.868 (± 0.049)	0.857 (± 0.025)	0.862 ± 0.029	0.799 (± 0.033)
Bramoullé	0.826 (± 0.065)	**0.883** (± 0.033)	0.862 ± 0.047	0.779 (± 0.051)
Manski	**0.882** (± 0.059)	0.848 (± 0.049)	**0.864** (± 0.037)	**0.805** (± 0.040)

Les trois modèles démontrent une capacité de prédiction satisfaisante pour ce jeu de données (près de 80% de justesse pour les trois modèles). Les résultats suggèrent une absence d'effet de pairs puisque le modèle de Bramoullé a une moins bonne justesse que les modèles CP et Manski, mais de meilleures valeurs de rappel et f-mesure moyennes même si les différences ne sont pas complètement significatives, compte tenu de la valeur des écarts types. Le modèle de Manski obtient la précision moyenne la plus élevée, indiquant une meilleure capacité à éviter les faux positifs pour la spécialité *généraliste*.

5.3 Evaluation de l'approche à base de règles

A partir de l'ensemble R des règles obtenues par AMIE3 avec les paramètres $s\_hc = 1\%$, un nombre de littéraux max (tête incluse) de 4, et le biais B_1, puis sélectionnées par le biais B_2, nous avons implémenté le classifieur suivant : on applique la règle de plus forte confiance, puis de plus fort support si plusieurs règles ont la même confiance. Si aucune règle ne se déclenche, on assigne à l'exemple la classe minoritaire (*généraliste*). Ceci est motivé par le fait que si aucune des nombreuses règles concluant sur *autre* ne se déclenche, on préfère assigner à la classe *généraliste*, modélisant ainsi le fait que le choix *généraliste* est un choix par défaut.

Le tableau 4 présente le nombre de règles générées R_{pairs} pour différents seuils de confiance s_{pca} en distinguant les règles sociales Rso (i.e. comportant au moins un prédicat Ami_de) des règles simples Rsi, et en précisant pour chacun de ces ensembles le nombre de règles concluant sur *généraliste* (Rsi_Gen et Rso_Gen). Ces statistiques montrent qu'en moyenne les règles sociales représentent 32,62 % des règles générées et que quel que soit le seuil s_{pca}, l'approche découvre peu de règles pour *généraliste*. Le tableau 4 montre également que les biais spécifiques définis (B_2) permettent de supprimer de 30% à près de 50% des règles (valeur de R_{supp}), ce qui montre leur intérêt.

TAB. 4 – *Nombre de règles en moyenne de chaque type de règles (10-validation croisée, les chiffres donnés sont des moyennes sur les 10 plis).*

s_{pca}	$\|R_{pairs}\|$	$\|Rsi\|$	$\|Rsi\_Gen\|$	$\|Rso\|$	$\|Rso\_Gen\|$	$\|R_{supp}\|$
1	478.2	357,8	3.2	120,4	0,6	144,3 (30,17%)
0,9	769,6	528,6	5,9	241	0,6	298,4 (38,74%)
0,8	1195	700,0	32,9	495	1,6	597,2 (49,93 %)

Le tableau 5 présente les mesures classiques de précision, rappel, F-mesure et justesse, et montre que la meilleure justesse est obtenue pour une valeur de s_{pca} de 1 (0.789). Les performances en terme de justesse sont comparables à celles obtenues par le modèle linéaire en moyenne. De plus, les résultats obtenus par cette deuxième approche confirme l'absence d'effet de pairs. En effet, si les règles faisant intervenir une ou plusieurs relations sociales sont ajoutées aux règles simples, la justesse baisse significativement, en particulier pour des valeurs de s_{pca} faible. Comme l'approche génère de nombreuses règles concluant sur la spécialité *autre* comparé au nombre de règles pouvant expliquer une préférence pour *généraliste*, l'approche est beaucoup moins efficace dans ce cas que le modèle statistique. Le rappel (pour la classe *généraliste*) est en particulier beaucoup plus faible quel que soit l'ensemble de règles considéré. Les performances baissent en même temps que le seuil de confiance pour une même valeur s_{hc} du seuil de couverture. En effet, des règles de faible confiance qui classent des exemples comme *autre* apparaissent, et la règle par défaut qui classe comme généraliste ne s'applique plus.

5.4 Discussion

Les deux approches (statistique et à base de règles) suggèrent l'absence d'effet de pairs compte tenu de l'absence d'impact significatif des variables et littéraux sociaux sur les prédictions effectuées. Les modèles ont des performances similaires mais se différencient par leur capacité à prédire les préférences pour la spécialité généraliste.

TAB. 5 – *Performances en 10-validation croisée pour les modèles de règles (moyenne (± écart-type) sur les 10 plis*

s_{pca} - R	Précision	Rappel	F-mesure	Justesse
1 - Rsi	0.618 (± 0.083)	**0.49** (± 0.163)	**0.536** (± 0.129)	**0.789** (± 0.037)
1 - R_{pairs}	0.608 (±0.124)	0.364 (± 0.349)	0.443 (± 0.167)	0.777 (± 0.037)
0,9 - Rsi	0.649 (± 0.120)	0.388 (± 0.144)	0.479 (± 0.136)	0.787 (± 0.045)
0,9 - R_{pairs}	0.621 (± 0.423)	0.133 (± 0.112)	0.211 (± 0.169)	0.761 (± 0.035)
0,8 - Rsi	**0.708** (± 0.190)	0.264 (± 0.121)	0.373 (±0.141)	0.777 (± 0.041)
0,8 - R_{pairs}	0.492 (± 0.429)	0.087 (± 0.087)	0.142 (± 0.142)	0.751 (± 0.028)

Une différence importante est que si les modèles statistiques permettent de caractériser la significativité des variables explicatives en présence de toutes les variables sélectionnées, les modèles à base de règles permettent d'isoler des sous-populations qui ont un comportement spécifique. Ainsi une règle de confiance 1 : $Factspe\_medtech(?a, 2), Factspe\_precaires(?a, 3), Factspe\_soins\_spe(?a, 1) \rightarrow Best\_spe1(?a, generaliste)$ permet de conclure sur le choix de généraliste en considérant uniquement leur faible appétence pour les soins spécialisés, les actes techniques et une envie peu affirmée de s'occuper des précaires. De même, la règle sociale sûre suivante : $Factspe\_medtech(?a, 5), Reussite\_1ere\_ann\_med(?h, Reussit), Ami\_de(?a, ?h) \rightarrow Best\_spe1(?a, autre)$ prédit un choix différent de généraliste pour des étudiants sur la seule base de leur fort intérêt pour les actes techniques et la présence dans leur entourage d'au moins un étudiant ayant pu réussir sa première année du premier coup. Enfin, les performances du modèle de Bramoullé, et du modèle de règles comprenant les règles sociales pourraient être sous-estimées : la suppression des amis cités par certains étudiants, mais qui n'ont pas répondu au questionnaire, a diminué la richesse des relations sociales exploitables, limitant la capacité de ces modèles à capturer les effets des réseaux d'amitiés.

6 Conclusion

Nous avons présenté dans cet article une première démarche permettant de comparer deux types d'approches prédictives interprétables pour une étude de l'effet de pairs : un modèle linéaire en moyenne et une approche de fouille de règles pour laquelle des biais spécifiques ont été définis. Les expérimentations menées sur un jeu de données réelles semblent montrer que les deux approches sont cohérentes concernant l'absence d'un effet de pairs mais qu'elles apportent des informations différentes sur les variables explicatives qui peuvent être considérées et les sous-populations concernées. Dans de futurs travaux, nous comptons explorer les données de manière plus approfondie (autres spécialités, k meilleur choix de spécialité, autres biais), analyser d'autres jeux de données ou les effets de pairs sont avérés et combiner les deux types d'approches.

7 Acknowledgements

Ce travail a été effectué dans le cadre de l'ANR ERMES N°-23-CE36-0009.

Références

Blockeel, H. et L. D. Raedt (1998). Top-down induction of first-order logical decision trees. *Artif. Intell. 101*(1-2), 285–297.

Bramoullé, Y., H. Djebbari, et B. Fortin (2009). Identification of peer effects through social networks. *Journal of Econometrics 150*(1), 41–55.

Bramoullé, Y., H. Djebbari, et B. Fortin (2020). Peer effects in networks : A survey. *Annual Review of Economics 12*(Volume 12, 2020), 603–629.

Cropper, A. et S. Dumancic (2022). Inductive logic programming at 30 : A new introduction. *J. Artif. Intell. Res. 74*, 765–850.

Hanushek, E., J. Kain, J. Markman, et S. Rivkin (2003). Does peer ability affect student achievement ? *J. Applied Economics 18*, 527–44.

Huang, X., Y. Izza, A. Ignatiev, et J. Marques-Silva (2021). On efficiently explaining graph-based classifiers. In *Proceedings of KR'21*, pp. 356–367.

Lajus, J., L. Galárraga, et F. M. Suchanek (2020). Fast and exact rule mining with AMIE 3. In *The Semantic Web - 17th International Conference, ESWC 2020*, Volume 12123 of *Lecture Notes in Computer Science*, pp. 36–52. Springer.

Manski, C. F. (1993). Identification of Endogenous Social Effects : The Reflection Problem. *The Review of Economic Studies 60*(3), 531–542.

Meilicke, C., M. W. Chekol, D. Ruffinelli, et H. Stuckenschmidt (2019). An introduction to anyburl. In *KI 2019 : Advances in Artificial Intelligence - 42nd German Conference on AI, Germany, 2019*, Volume 11793 of *LNCS*, pp. 244–248. Springer.

Nassiri, A. K., N. Pernelle, et F. Saïs (2023). REGNUM : generating logical rules with numerical predicates in knowledge graphs. In *Proceedings ESWC 2023*, Volume 13870 of *LNCS*, pp. 139–155. Springer.

Ortona, S., V. V. Meduri, et P. Papotti (2018). Rudik : Rule discovery in knowledge bases. *Proc. VLDB Endow. 11*(12), 1946–1949.

Simonne, L., N. Pernelle, F. Saïs, et R. Thomopoulos (2021). Differential causal rules mining in knowledge graphs. In *Proceedings. K-CAP '21*, pp. 105–112. ACM.

Summary

Many econometric works have attempted to identify a peer effect in a population. In this paper, we propose to compare the results obtained by two types of interpretable predictive approaches: (1) an averaging linear model which assumes that individuals are affected in a linear fashion by the actions and mean characteristics of their peers and (2) a generalist first-order logic rule mining approach that has been adapted through the use of specific language biases. The experiments were carried out on a real data set concerning the choice of a medical specialty by medical students when they have to select their internship position.