UNIVERSITÉ DU QUÉBEC À MONTRÉAL

# APPLICATION D'UN MODÈLE DE PRÉVISION DE LA SANTÉ DES ENFANTS EN RELATION AVEC LES CONDITIONS ENVIRONNEMENTALES

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN ÉCONOMIQUE

PAR

ANTOINE THOMPSON-LEDUC

FÉVRIER 2018

La présente recherche a été menée grâce à un soutien financier accordé au Réseau canadien des centres de données de recherche (RCCDR) par le Conseil de recherches en sciences humaines (CRSH), les instituts de recherche en santé du Canada (IRSC), la Fondation canadienne pour l'innovation (FCI) et Statistique Canada, ainsi que la Chaire de recherche Industrielle Alliance sur les enjeux économiques et changements démographiques (CEDIA). Bien que les recherches et les analyses aient été faites à partir des données de Statistique Canada, les opinions exprimées ne représentent pas celles de Statistique Canada.

# REMERCIEMENTS

J'aimerais remercier mon directeur Pierre Ouellette et mon co-directeur Philip Merrigan pour leur soutien et leur approche personnalisée. Vous avez été une constante source de motivation et vous m'avez donné la chance de pousser mes limites.

J'aimerais remercier la Chaire de recherche Industrielle Alliance sur les enjeux économiques des changements démographiques (CEDIA), l'École des Sciences de la Gestion (ESG-UQAM), le centre interuniversitaire québécois de statistiques sociales (CIQSS) et mes parents pour leur soutien financier tout au long de ma maîtrise.

J'aimerais remercier mes collègues et amis de l'UQÀM et du CIQSS pour leur support technique et administratif durant cette recherche. Je pense à Frédéric Brousseau (CIQSS), Martine Boisselle-Lessard (UQÀM), Catherine Blanchard-Gougeon (CIQSS), mais aussi à tous ceux et celles au cinquième étage de l'ESG.

Finalement, j'ai la chance d'être soutenu par une famille exceptionnelle et des amis résilients. Vous êtes la raison pour laquelle je me suis embarqué dans cette aventure et pour laquelle je suis aussi fier d'en sortir avec succès.

TABLE DES MATIÈRES

# LISTE DES FIGURES

# LISTE DES TABLEAUX

# LISTE DES ABRÉVIATIONS, CYCLES ET ACRONYMES

| | |
|---|---|
| AEDC | *Australian Early Development Census* |
| AUC | *Area Under the Curve* |
| Ppb | Partie par milliard (*Parts per Billion*) |
| BDPQA | Base de données pancanadienne sur la qualité de l'air |
| CAS | Cote air santé |
| Coll. | Collaborateurs |
| D.É.S | Diplôme d'études secondaires |
| ELNEJ | Enquête longitudinale nationale sur les enfants et les jeunes |
| HRS | *Health and Retirement Survey* (États-Unis) |
| PCM | Personne ayant la meilleure connaissance de l'enfant |
| ROC | *Receiver Operating Characteristic Curve* |
| RNSPA | Réseau national de surveillance de la pollution atmosphérique |
| SSÉ | Statut Socioéconomique |
| $\mu g/m^3$ | Microgrammes (un millionième de gramme) par mètre cube |

# RÉSUMÉ

La détermination des risques de santé à l'enfance implique une tâche prédictive importante. Identifier ces facteurs de risques permet d'intervenir rapidement et de réduire les conséquences néfastes d'une mauvaise santé de l'enfant sur sa vie future. Bien que plusieurs déterminants socioéconomiques aient déjà été montrés comme causes importantes de mauvaise santé chez l'enfant, peu d'études utilisent des variables environnementales dans leur modèle.

C'est en ce sens que ce mémoire s'intéresse à l'apport de variables environnementales dans la prédiction de la santé de l'enfant. Les données utilisées proviennent de l'échantillon d'enfants longitudinal de l'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ) et de la Base de données pancanadienne sur la qualité de l'air (BDPQA). Les prédictions sont faites avec différents modèles de régressions et de méthodes en arbre à plusieurs stades de l'enfance, avec et sans variables environnementales.

Les résultats révèlent que les deux meilleures méthodes de prédictions sont la méthode lasso avec forme fonctionnelle Logit, ainsi que la méthode *boosting*. Les trois variables les plus prédictives de la santé de l'enfant sont la santé de la mère à la naissance, la santé de l'enfant à la naissance et le revenu familial. L'ajout de variables environnementales dans la prédiction semble surspécifier les modèles et ainsi, nuire à leur performance.

MOTS-CLÉS : santé, enfant, modèles de prédiction, Logit, Probit, apprentissage automatique, ELNEJ, BDPQA

# INTRODUCTION

À l'enfance, les processus de propagation de santé sont analysés dans de nombreux domaines de recherche, comme la psychologie, l'épidémiologie et la santé publique. Identifier les facteurs de risques de la santé de l'enfant permet d'intervenir rapidement et de réduire les conséquences néfastes sur sa vie future. D'un point de vue économique, l'investissement en santé à un stade précoce du développement aura, via ses effets multiplicateurs, un impact positif sur sa santé future, sa réussite scolaire, et son développement cognitif (Heckman, 2006). L'identification des enfants à risque d'être en mauvaise santé implique une tâche prédictive importante, étroitement liée à la question causale.

La relation entre le statut socioéconomique et la santé compte parmi les relations les plus robustes et les plus documentées en sciences sociales. Parmi les déterminants généralement inclus dans les recherches sur la santé de l'enfant, on retrouve l'éducation parentale, le salaire des parents et la situation familiale. Bien que le lien entre la qualité de l'air et la santé soit de plus en plus documenté et robuste, peu d'études incluent des variables de pollution. Le but de cette recherche est de proposer un modèle de prédiction de la santé d'un enfant qui inclut des variables environnementales.

Ce mémoire consiste à prédire la santé de l'enfant à l'âge de quatre, six, huit et dix ans, à partir d'information périnatale. Des variables sur la qualité de l'air seront ajoutées à un modèle de prédiction, afin de mesurer l'apport de la pollution de l'air lors de la grossesse dans les mesures de performances du modèle prédictif. Aussi, des modèles d'apprentissage automatique seront appliqués, afin d'améliorer les prédictions.

L'échantillon est une cohorte de 2433 enfants suivis de la naissance à l'âge de 10 ans provenant de l'enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ)

produite par Statistique Canada. Au premier cycle de l'enquête (1994), l'information récoltée sur les nouveau-nés comporte des variables sur la grossesse de la mère, le statut socioéconomique des parents, la dynamique familiale et des caractéristiques physiques à la naissance de l'enfant. Durant les quatre cycles biennaux suivants, une variable de santé de l'enfant, sur une échelle de 1 à 5, sera transformée en variable dichotomique de mauvaise santé et servira de variable dépendante du modèle. Les variables environnementales du modèle prédictif de la recherche proviennent de la base de données pancanadienne sur la qualité de l'air (BDPQA) qui récolte la mesure des principaux polluants environnementaux à travers le Canada pendant la période étudiée.

Les modèles de prédiction binomiale de type Probit et Logit seront tout d'abord appliqués, avec et sans variables environnementales. Ces méthodes seront comparées avec des méthodes d'apprentissage automatique, permettant de proposer des formes d'estimation plus flexibles et des modèles de régularisation. Les méthodes de régularisations sont utilisées pour laisser à l'algorithme le choix (la détermination) des variables offrant un pouvoir explicatif, cela permettra de confirmer, ou non, le choix de variables fait à partir d'intuitions économiques des chercheurs sur les facteurs de risque de la santé de l'enfant. De plus, ces modèles seront comparés à des modèles en arbre par itération (forêts aléatoires). Les modèles en arbre sont connus pour être facilement interprétables, de bien performer lorsqu'en présence de plusieurs effets d'interactions et, avec itérations, de bien prédire lors de l'ajout d'itérations.

Ce mémoire débutera par une revue de littérature sur les différents déterminants socioéconomiques de la santé de l'enfant, du lien entre la qualité de l'air et la santé de l'enfant et des méthodes de prédiction utilisées. Le deuxième chapitre présentera les méthodes utilisées dans le mémoire. Le troisième chapitre présentera les résultats des méthodes appliquées. Le quatrième chapitre comporte une discussion sur les résultats obtenus. Finalement, le travail terminera par une conclusion.

# CHAPITRE I

# REVUE DE LITTÉRATURE

## 1.1 Déterminants socioéconomiques de la santé de l'enfant

Les déterminants de la santé de l'enfant ont été étudiés dans plusieurs études causales basées sur des modèles économiques de propagation de santé. Parmi ces études, le travail de Case, Lubotsky et Paxson (2002) est souvent un point de départ en recherche. Les auteurs étudient l'impact du revenu familial sur la fréquence et la longévité de chocs de mauvaise santé. L'échantillon provient du *National Health Interview Survey* (NHIS)[1], qui regroupe plus de 180 000 observations d'enfants et de jeunes aux États-Unis de 0 à 17 ans entre 1984 à 1995. En séparant l'échantillon par groupe d'âge (0-3 ans, 4-8 ans, 9-12 ans et 13-17 ans), une régression par Probit ordonnée évalue la santé d'un enfant en relation avec le revenu, l'éducation parentale et des variables de contrôle de caractéristiques familiales. La santé de l'enfant est évaluée par la mère, sur une échelle de 1 à 5 (1 = excellente santé et 5 = mauvaise santé). Les résultats montrent que le prédicteur principal est le revenu familial (coefficient de -0,183). De plus, les impacts du niveau de revenu familial de long terme semblent augmenter avec l'âge (doubler le revenu familial augmente la probabilité d'être en bonne ou excellente santé de 4 pour cent chez les 0-3 ans, 4,9 pour cent chez les 4-8 ans, 5,9 pour cent chez les 9-12 ans et 7,2 pour cent chez les 13-17 ans). Le comportement de la mère pendant la grossesse

---

[1] Deux extensions au NHIS sont utilisées ; l'extension de la santé des jeunes en 1988 (*Child Health*. NHIS-CH) et l'étude en données panel du revenu de 1997 (*Child development supplement* PSID-CDS).

(par exemple, l'alcool ou le tabac) ou la situation socioéconomique des parents (par exemple, le salaire familial ou l'éducation parentale) ont des effets significatifs sur la santé de l'enfant.

Dans le contexte canadien, il est possible que l'assurance maladie publique amoindrisse l'effet du revenu. Cependant, ces résultats semblent robustes à une transposition dans le contexte canadien (Currie et Stabile, 2003 et Allin et Stabile, 2012). Currie et Stabile (2003) utilisent les observations de 10 000 enfants de 0 à 11 ans de l'ELNEJ, suivis de 1994 à 1998. La santé ($S$) de l'enfant $i$ au temps $t$ est mesurée sur une échelle d'un à cinq par la personne ayant la meilleure connaissance de l'enfant *PCM* (92 % des cas, la mère). La régression du modèle Probit est basée sur les variables suivantes; le logarithme du revenu familial (dollars réels canadiens de 1997), l'éducation de la mère, un ensemble de variables temporelles et un ensemble de variables de santé de l'enfant à la naissance. Les résultats principaux sont que la probabilité d'avoir une mauvaise santé s'accroît significativement avec un statut socioéconomique (SSÉ) faible. Le coefficient du revenu décroît de 0,140 entre le groupe des « 0 à 3 ans » par rapport aux « 13 à 15 ans », contre 0,120 pour Case et coll. (2002).

Une variété de mécanismes peut expliquer une relation positive entre le revenu et la santé de l'enfant. On peut penser qu'un revenu élevé peut allouer plus de ressources à des soins de santé de meilleure qualité, ou indirectement, un enfant avec des parents plus productifs sur le marché du travail adopte un mode de vie plus sain, ce qui influe la santé de l'enfant.

Le revenu familial n'est qu'un parmi plusieurs prédicteurs de la santé de l'enfant. Dans les deux études, le coefficient du revenu chute de près d'un tiers lorsque l'on ajoute l'éducation de la mère (passe de -0,183 à -0,114 dans l'étude Case et coll. et de -0,151 à -0,132 dans celle de Currie et Stabile). L'effet de l'éducation est capté par le revenu lorsque celui-ci n'est pas inclus dans la régression. Cela peut être expliqué par la corrélation forte entre le revenu familial et l'éducation parentale. Une personne

éduquée est plus informée sur la manière d'adopter un mode de vie sain (via l'alimentation et le type d'exercice). Ces résultats ont été étudiés pleinement dans plusieurs articles (Feldman, 1989; Cutler et Lleras-Muney, 2006; et Currie et Moretti, 2002).

Parmi les autres déterminants de la santé durant l'enfance, la santé de l'enfant observée en bas âge est un déterminant significatif sur les états de santé futurs (Case, Fertif et Paxson, 2004; Contonyannis et Li, 2011). L'analyse de Contoyannis et Li (2011) étudie la persistance des problèmes de santé entre l'enfance et l'adolescence en utilisant la base de données de l'ELNEJ avec une régression Probit ordonnée. La mesure de santé à la période précédente ($t$-1) est un estimateur significatif de la santé à la période observée (coefficient de 0,752, écart type de 0,022). Avec l'ajout d'effets fixes, le coefficient reste positif et significatif, mais nettement moins élevé (0,296 avec un écart-type de 0,027). L'impact des différences stables et inobservées entre les personnes est donc capté par les coefficients si on ne rajoute pas d'effets fixes. L'hétérogénéité inobservée (effets aléatoires) permet une amélioration de prévision de 34 %.

Durant la période prénatale, la santé de l'enfant à la naissance sera influencée par la nutrition et la santé de la mère, ainsi que de l'aide médicale que la mère reçoit durant sa grossesse. Case, Fertif et Paxson (2004) trouvent une relation causale de l'utilisation du tabac pendant la grossesse avec le développement de problèmes cognitifs et une mauvaise éducation. La consommation d'alcool durant la grossesse influence le poids de l'enfant à la naissance, le développement neuronal, ainsi que les anomalies faciales à la naissance (Case et Paxson, 2002).

## 1.2 Santé de l'enfant et les variables environnementales

La qualité de l'air et ses impacts sur la santé sont de plus en plus étudiés dans la littérature scientifique. Les enfants font partie des groupes les plus susceptibles d'être affectés par l'environnement. L'exposition à la pollution de l'air lors de la grossesse provoque des effets négatifs sur la santé de l'enfant. L'exposition à l'ozone ($O_3$) et au monoxyde de carbone (CO) augmente les chances de problèmes de petit poids de l'enfant à la naissance et de retard de croissance (Salam et coll., 2005).

L'étude de Burnet et coll. (2008) utilise les données d'admissions hospitalières et de pollution dans cinq centres urbains australiens entre 1998 et 2001[2] pour estimer l'impact de différents polluants sur l'admission hospitalière. Les polluants considérés sont les matières particulaires ($PM_{2,5}$ et $PM_{10}$)[3] , le monoxyde de carbone, l'ozone, le dioxyde de soufre et le dioxyde d'azote. À partir d'une étude croisée désaisonnalisée en contrôlant pour la température, les valeurs extrêmes de chaleur et les jours fériés, les auteurs trouvent des impacts significatifs entre les différents polluants et la fréquentation hospitalière pour les pneumonies et les bronchites aiguës (0 à 4 ans), les problèmes respiratoires (tous les âges) et l'asthme (« 5-14 ans »). La relation la plus importante trouvée est une augmentation de six pour cent d'admission hospitalière chez les « 5-14 ans » en relation avec une augmentation de 5,1 *ppb*[4] de dioxyde d'azote ($NO_2$) en 24 heures.

---

[2] Centres urbains : Brisbane, Canberra, Melbourne, Perth et Sydney.
[3] Les indices des matières particulières et la taille du diamètre en micromètre.
[4] Ppb : partie par milliard (*Parts per billion*)

# CHAPITRE II

# MÉTHODOLOGIE

## 2.1 La causalité *vs* les prédictions

Les analyses statistiques ont plusieurs objectifs, parmi lesquels on retrouve le développement de modèles de causalité et ceux de prédictions. Jusqu'à présent, les études présentées dans cette revue de littérature n'étaient que causales. Dans le domaine économique, les modèles statistiques sont presque exclusivement utilisés à des fins causales et les modèles ayant de forts pouvoirs explicatifs sont censés être de bons prédicteurs. Pourtant, les deux types de modèles jouent un rôle dans l'élaboration et l'évaluation de certaines théories et politiques.

Les recherches économiques de causalité sont basées sur le travail collaboratif d'économistes et d'experts en santé, ainsi que sur la littérature scientifique. Les hypothèses de recherche sont amenées par construction et non par mesures. C'est avec ces hypothèses que les auteurs construisent un modèle statistique et choisissent les variables à inclure dans le modèle. Les conclusions statistiques sont transposées en conclusion de recherche et souvent transformées en recommandations politiques. Ces modèles explicatifs sont présentés comme une application d'un modèle statistique de données pour tester une hypothèse causale ou d'expliquer un phénomène à partir de construction théorique.

Les modèles de prédictions cherchent plutôt à appliquer un modèle statistique ou un algorithme dans le but de prédire les réalisations futures de variables d'intérêt. Cette définition implique souvent de la prédiction temporelle (évaluer $Y$ au temps $t + k$, $k > 0$, à partir d'observations au temps $t$). Dans le cas d'exercices de prédiction, on évalue la

performance des modèles avec des observations qui n'ont jamais été considérés pour l'estimation du modèle. Pour se faire, les observations sont divisées aléatoirement en deux groupes. Les paramètres sont estimés avec l'échantillon d'apprentissage (*training set*) et leur performance sera testée sur l'échantillon de test (*test set*).

Dans les études causales, les mesures fréquemment utilisées pour mesurer le pouvoir prédictif sont habituellement le $R^2$ et la significativité de certains paramètres obtenue avec la valeur-*F*. En se basant sur ces mesures, ajouter de l'information ou des variables indépendantes ne fera qu'améliorer la performance du modèle. La performance des modèles causaux peut être surévaluée, car cette technique ne considère pas l'erreur inhérente au modèle par rapport à la vraie relation entre les variables indépendantes et la variable dépendante. Le biais du modèle sera minimisé, mais la variance restera trop grande. C'est ce que l'on appelle des problèmes de surspécification (traduit de l'anglais «*overfitting*») (James, Hastie, Tibshirani et Witten, 2014). Dans le monde réel, l'éligibilité des programmes d'intervention de santé est souvent basée sur les facteurs de risques qui prédisent bien la santé de l'enfant.

De plus, les nouvelles bases de données sont de plus en plus complexes et les relations entre les différentes variables sont difficiles à établir. Les nouveaux outils économétriques, comme l'apprentissage automatique (*Machine learning*), sont des méthodes qui performent mieux que les régressions dans un contexte de prédiction (Mullainathan et Spiess, 2017). Les modèles d'apprentissage automatique en arbre sont les meilleurs pour capter les effets d'interaction entre les différentes variables. En conséquence, les modèles de prédictions peuvent améliorer les modèles explicatifs (Schmuelli, 2010).

En sciences économiques, il est nécessaire de faire des recherches avec des méthodes de prédictions. Les modèles de prédiction permettent d'établir la distance entre la théorie et la pratique, de tester les modèles sur des données qui n'ont jamais été utilisés, ainsi que de traiter de grandes bases de données en comprenant des relations

hétérogènes et complexes entre les variables. Étant donné que l'on ne connait pas la relation exacte entre les variables indépendantes $X$ et la variable dépendante $Y$, les modèles de prédictions hors échantillon donnent une référence de performance. Si un modèle causal est loin d'atteindre la référence des performances d'un modèle de prédiction, cela met la table pour une amélioration pratique et théorique des modèles étudiés. Cela dit, souligner la relation causale des variables permet d'établir une base théorique de variables de santé à utiliser.

Il est difficile d'isoler les effets de la pollution de l'air sur la santé, car il y a de fortes corrélations entre le statut socioéconomique et l'emplacement géographique des enfants. Par exemple, les enfants moins aisés auront tendance à se situer dans les quartiers plus industriels et, de ce fait, plus pollués. L'endogénéité de la propagation de la santé implique des hypothèses économétriques lourdes et, parfois, non testables. L'inférence statistique peut passer par la démonstration que la pollution de l'air joue un rôle significatif dans la prédiction de la santé de l'enfant. En somme, les modèles de prédictions sont un ajout justifié à la littérature scientifique dans ce domaine.

L'étude de Chittleborough *et al.* (2010) utilise un modèle de prédiction dans le but de confirmer ou non si les facteurs de risques utilisés dans les études causales sont de bons prédicteurs sur la santé de l'enfant à l'aide de variables périnatales. En utilisant les données d'une enquête australienne sur le développement de l'enfant (AEDC), les chercheurs procèdent un modèle de régression log-poisson sur 13 000 enfants et prédisent les problèmes de l'enfant à l'âge de cinq ans en utilisant 22 prédicteurs et en mesurant la performance des modèles avec le critère AUC (*Area Under the Curve,* voir section 3.2). Les chercheurs montrent qu'un modèle contenant six variables périnatales performe aussi bien qu'un modèle qui en contient 22 (les variables retenues sont l'âge maternel, fumer durant la grossesse, le statut marital et l'emploi des deux parents). Le AUC des modèles est de 0,682 pour les garçons et 0,724 pour les filles. Avec ces modèles, approximativement dix pour cent des enfants nécessiteraient des soins

intensifiés à la naissance et, en implantant un programme d'intervention, un quart des problèmes de développement pourraient être prévenus à l'âge de cinq ans.

La recherche de Rooney, Mathason et Schauberger (2010) cherche à prédire l'obésité infantile à partir de caractéristiques durant la grossesse et la naissance. Les données proviennent de dossiers médicaux de mille mères du Midwest américain suivies pendant une période de 10 à 15 ans. Les variables incluent le sexe de l'enfant, la santé de la mère, les assurances de santé et le type d'accouchement. Le modèle logistique multivarié est construit à partir des variables significatives des modèles univariés. Le modèle final a ensuite été soumis à une méthode de sélection à rebours (*Backward Selection*). Les résultats montrent que l'obésité maternelle est le meilleur prédicteur de l'obésité chez l'enfant. De plus, le risque d'obésité infantile double lorsque l'enfant, à la naissance, pèse plus de 8,5 livres. Les autres facteurs de risques significatifs sont les accès aux assurances maladie privées et les enfants nés par césarienne. Les résultats de ces deux recherches de prédiction sont en accord avec celles des recherches causales.

## 2.2 Méthodes utilisées[5]

Les méthodes visitées dans ce mémoire sont des méthodes de prédiction en classification binaire. La variable dépendante $y$ prend la valeur 0 s'il n'y a pas de problèmes de santé et $y = 1$ sinon. Réduire le nombre de catégories (de 5 à 2) permet d'augmenter le nombre d'observations par catégorie et de réduire la variance reliée à la subjectivité du PCM lors de la réponse.

$$y_i = \begin{cases} 0, & \text{avec une probabilité } \quad p \\ 1, & \text{avec une probabilité } (1-p) \end{cases}$$

Il existe plusieurs façons d'estimer ces probabilités. Il convient d'étudier plusieurs méthodes d'estimation, afin de trouver celle qui performe le mieux pour la prédiction.

Intuitivement, le modèle optimal est celui qui minimise les erreurs de classification (*EC*). Cette mesure est déterminée par le pourcentage des prédictions $\hat{y}_i$ qui ne correspond pas à la valeur observée de $y_i$ .

$$EC = \frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_i \neq y_i)$$

Pour minimiser les erreurs de classification, il est optimal d'utiliser le critère bayésien pour classifier les prédictions. Par conséquent, si les résultats des méthodes prédisent une probabilité au-delà de 0,5 pour une observation, la prédiction sera de 1.

---

[5] Les auteurs de références utilisés sont Stock et Watson, dans *Principes d'Économétrie*, au chapitre de modèles à variables dépendantes binaires (chapitre 11), ainsi que Wooldridge, dans *Approche Moderne d'économétrie*, au chapitre d'effets linéaires dans les modèles longitudinaux (chapitre 10) et au chapitre des modèles à réponses binaires (chapitre 15). Les modèles en apprentissage automatique proviennent du livre *Introduction to Statistical Learning* de Hastie et James (2014).

Étant donné qu'il est possible d'avoir un problème de classe faible (exemple : la mauvaise santé de l'enfant n'est observée que dans 12 % des cas), il est préférable d'utiliser la mesure de sensibilité et de spécification pour discriminer par rapport au type d'erreur commis. Ces mesures sont obtenues à l'aide de la matrice de confusion.

TABLEAU 2.1 : Matrice de confusion

| | | Vraies Valeurs | |
|---|---|---|---|
| | | $y_i = 0$ | $y_i = 1$ |
| Prédictions | $\hat{y}_i = 0$ | Vrai Négatif (VN) | Faux Négatif (FN) |
| | $\hat{y}_i = 1$ | Faux Positif (FP) | Vrai Positif (VP) |

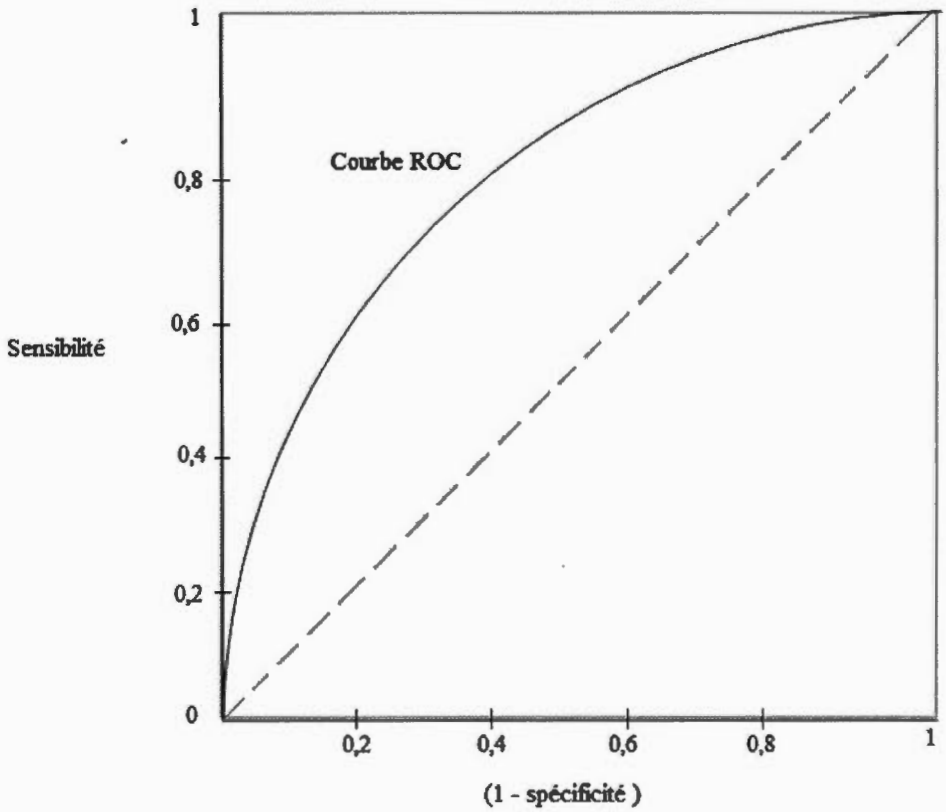
La sensibilité du modèle, ou le taux de vrais positifs, mesure la proportion de prédiction positive ($\hat{y}_i = 1$) qui identifie correctement les vraies valeurs (le pourcentage d'enfants en mauvaise santé correctement identifiés comme étant en mauvaise santé). La spécificité, ou le taux vrais négatifs, mesure la proportion de prédiction négative ($\hat{y}_i = 0$) correctement identifié (le pourcentage d'enfants en santé qui sont correctement identifiés comme étant en santé). En suivant la matrice de confusion, la sensibilité du modèle est exprimée par $(VP/(VP + FN))$ et la spécificité $(VN/(VN + FP))$. Ces types d'erreurs sont comparables aux erreurs de type 1 pour (1-spécificité) ainsi que les erreurs de type 2 pour (1-sensibilité)

Si la sensibilité d'un modèle est trop basse, une grande proportion d'enfants avec une mauvaise santé passera sans alerte et ne pourra se classer dans un programme d'intervention. D'autre part, si la spécificité est trop large, nous pourrions cibler trop d'enfants en bonne santé et cela serait inefficace. La sensibilité et la spécificité dépendent des différents points de rupture (*cutoff points*) utilisés et l'arbitrage entre ces deux mesures est essentiel. Les points de ruptures possibles varient entre 0 et 1 et servent à classer les prédictions par rapport aux probabilités prédites des modèles. Le

critère bayésien propose la rupture à 0,5, cela fait en sorte que la classification sera de 1 si la probabilité prédite est supérieure à 0,5. Bien que le critère bayésien minimise les erreurs de classification, une distribution inégale des différentes classes peut encourager la décision du point de rupture de diverger du critère bayésien.

Un moyen de mesurer la performance du modèle en opérant un arbitrage entre ces deux mesures est d'utiliser la courbe ROC (*Receiver Operating Characteristic*). La ROC est un outil graphique qui représente, pour chaque point de rupture (*cutoff*) un point sur le graphique. On retrouve sur l'axe des abscisses du graphique la spécificité et sur l'axe des ordonnées la sensibilité.

FIGURE 2.1 : Représentation graphique de la courbe ROC



L'avantage d'utiliser la courbe ROC est que cela épargne le coût du choix du point de rupture (*cost-complexity*). Pour un point de rupture à 100 ($\hat{y}_i = 0 \; \forall \; i$), le taux de sensibilité est égal à 1, car il n'y a aucun faux négatif. Pour un point de rupture à 0 ($\hat{y}_i = 1 \; \forall \; i$), nous nous assurons qu'il n'y ait aucun faux positif, mais on ne classe aucun $y$ à un. Plus la courbe est concave, plus notre modèle est performant, car les taux de sensibilité et de spécificité sont élevés. Parfois, les courbes s'entrecroisent et il est difficile d'identifier le modèle le plus performant, c'est pourquoi on utilise l'outil numérique AUC (*Area Under the Curve*) pour mesurer empiriquement la performance du modèle. Un AUC de 0,5 signifie qu'un modèle ne fait pas mieux qu'une assignation aléatoire de $y$.

### 2.2.1 Méthodes de régression

En présence de variable dépendante binaire, l'espérance de l'état de santé d'une personne, tenant en compte l'information disponible en $t$, est égale à la probabilité que l'enfant développe une mauvaise santé, ainsi le modèle prédictif de régression permet une estimation de la probabilité que l'enfant soit en mauvaise santé à une période ultérieure à $t$, soit $t+k$.

$$E[Santé_{t+k} \mid X_t . X_{t-1} . X_{t-2} . \ldots . X_{t-T}] = P(Santé_{t+k} = 1 \mid X_t . X_{t-1} . X_{t-2} . \ldots . X_{t-T})$$

La technique des moindres carrés ordinaires (MCO) ne contraint pas les probabilités prédites à être bornées entre zéro et un. Un modèle bornant ces probabilités est approprié et est donné par la fonction de distribution F(.), une fonction cumulative. Le modèle logistique pour F(.) a l'avantage de fournir une forme fonctionnelle facile à manipuler. Elle s'écrit ainsi :

$$F(\beta' X_i) = \Lambda(\beta' X_i) = \frac{1}{1 + e^{-(\alpha + \beta X_i)}}.$$

Les coefficients de ces modèles sont estimés par la méthode du maximum de vraisemblance. Le modèle Probit s'écrit:

$$\Pr(Santé = 1) = \Phi(\alpha + \beta X_i),$$

où $\Phi$ est la fonction de répartition de la loi normale centrée réduite et $X_i$ est un vecteur de $K$ régresseurs.

### 2.2.2 Méthodes de régularisation

Il est habituel de choisir toutes les variables possibles pour former une prédiction. Par contre, inclure le maximum de variables possible risque d'entraîner un problème de surspécification (*overfitting*) lors des estimations sur l'échantillon de modelage entrainant des mauvaises prévisions dans l'échantillon test. Autrement dit, l'ajout de variables inutiles risque d'augmenter la variance des estimations sans réduire le biais. Pour régler ce problème, on utilise des méthodes d'élagage ou de régularisation pour réduire le nombre de variables ou de pénaliser l'ajout de coefficients dans le modèle prédictif. Les méthodes de régularisation ajoutent un terme de pénalité à la fonction objective pour réduire l'importance des régresseurs en poussant les coefficients vers zéro. Parmi les méthodes de régularisation, la méthode *lasso* sera utilisée dans ce mémoire.

Le *lasso* utilise le terme de pénalité qui fera pression sur les coefficients pour les rétrécir (*shrink*), ou les faire tendres vers zéro. Les coefficients du modèle de régression sont alors choisis en maximisant la fonction log-vraisemblance du Logit, sous la contrainte de pénalité du modèle, soit;

$$\sum_{i=1}^{N} Y_i \log\left(\Lambda(\beta' X_i)\right) + \sum_{i=1}^{N} (1 - Y_i) \log(\Lambda(\beta' X_i)) - \lambda \sum_{j=1}^{p} |\beta_j|$$

Cette maximisation sera, donc, soumise à deux pressions, soit celle de maximiser la vraisemblance (avec $p$ le nombre de variables explicatives sans la constante), qui aura tendance à augmenter lors de l'ajout de variables, et un terme de pénalité $\lambda \sum_{j=1}^{p} |\beta_j|$, qui encouragera le retrait de variables. La régression *lasso* ira jusqu'à écarter des coefficients du modèle (les pousser directement à zéro).

Le choix du terme $\lambda$ est très important. Une estimation avec un $\lambda$ élevé produira des coefficients près de zéro ou égaux à zéro. Le $\lambda$ optimal est celui qui performera le mieux sur l'échantillon test. Dans notre cas, puisque le nombre d'observations n'est pas très grand, nous déterminerons $\lambda$ par une méthode de validation croisée. En effet, un échantillon test est généralement 40% de l'échantillon, dans notre cas cela impliquerait une importante perte d'observations pour l'estimation. La validation croisée est un processus qui partitionne aléatoirement l'échantillon également en $k$ sous-groupes, appelés plis (*folds*). L'estimation implique d'estimer le modèle avec *k-1* plis à $k$ reprises. À chaque reprise, un pli différent n'est pas utilisé pour l'estimation, mais utilisé pour valider le modèle. C'est-à-dire que la mesure de performance est calculée avec le pli inutilisé dans l'estimation. Donc pour une valeur de $\lambda$, le modèle est estimé à $k$ reprises, et à chaque fois une mesure de performance est calculée. Ensuite, la moyenne de ces mesures est calculée une fois les $k$ estimations terminées. Cette procédure est répétée pour un grand nombre de valeurs de $\lambda$ (*grid search*). Le modèle retenu correspondant au $\lambda$ optimal qui est celui avec la meilleure performance. Si $k = n$, la technique de validation croisée $k$ est en fait du *bootstrap*.

### 2.2.3 Méthodes en arbre

Les principaux avantages des modèles en arbres sont leur interprétabilité et qu'ils ne rencontrent aucun problème d'assignation des interactions et des polynômes dans les

spécifications du modèle prédictif. Ce qui n'est pas le cas du *lasso*, car il faut spécifier précisément les effets d'interactions que l'on veut introduire dans le modèle.

2.2.3.1 La construction de l'arbre de régression.

Les méthodes d'arbres en classification suivent ce concept simple; on procède en une série de division binaire par une série de règles imposées. Chaque division binaire forme un nœud interne, qui entraînera deux branches. Les deux groupes seront divisés indépendamment l'un de l'autre, en deux autres groupes et ainsi de suite. Le processus est répété jusqu'à ce que les sous-groupes formés par l'arbre atteignent une règle d'arrêt. Un exemple serait d'arrêter lorsqu'il ne reste que cinq observations par groupes restants. Chaque variable peut être réutilisée dans les arbres, par exemple, il se peut que l'âge soit subdivisé une première fois, puis une deuxième fois dans un sous-groupe créé plus tard dans la construction de l'arbre. Chaque individu, suivant ces divisions, terminera dans un sous-groupe final, qu'on appelle un nœud terminal (*terminal node).* La prédiction de la classe pour l'individu sera la classe la plus fréquente dans le sous-groupe du nœud terminal. La question est alors de savoir comment former les nœuds internes (*split*) pour former les meilleures prédictions possibles.

La division binaire est déterminée selon deux critères; le seuil et l'ordonnancement des variables. Le seuil est utilisé pour déterminer la valeur de la variable qui divisera en deux l'échantillon. L'ordonnancement servira à choisir la variable finale à mettre dans l'arbre. Pour obtenir le seuil pour une variable $j$, il suffit de faire une prédiction de tous les seuils possibles pour une variable et d'utiliser le seuil qui réduit le plus possible l'erreur de classification du modèle. L'ordre des variables peut être obtenu en calculant, une fois le seuil $s$ pour chaque variable calculée, les modèles possibles avec les variables que l'on a sous la main. Une fois le premier nœud déterminé, nous refaisons ce processus pour les deux branches indépendamment.

Cependant, cette technique mène à des arbres compliqués, ininterprétables et suridentifiés. Une des solutions proposées pour alléger est méthode d'élagage de

l'arbre (*pruning*) une fois l'arbre construit. Ce principe cherche à pénaliser pour le nombre de nœuds de l'arbre en imposant un terme de pénalité. En prédictions, la variance des estimations d'un arbre unique est trop élevée. Pour corriger ce problème, il existe plusieurs méthodes d'estimation par itération dans le but de réduire la variance. La variance des estimations de ces arbres peut être réduite avec les méthodes d'itération comme le *bagging*, les forêts aléatoires et le *boosting*.

#### 2.2.3.2 Méthode à forêts aléatoires

La méthode avec forêts aléatoires estime un nombre $X$ d'arbres (sans élagage) et, pour chaque arbre, fait une prédiction $y$. Chaque arbre est estimé avec un échantillon bootstrap de l'échantillon initial. Dans les modèles de classification, la classe prédite pour chaque $i$ sera la classe la plus souvent prédite parmi les $X$ prédictions pour l'observation $i$. Dans notre cas, si la valeur est positive plus de 50 % du temps dans les échantillons bootstrap, la prédiction finale sera de 1. Lors de la formation des arbres, seulement un échantillon aléatoire de $m$ prédicteurs est choisi comme candidat pour déterminer le prochain nœud. Typiquement, le nombre de prédicteurs (noté $m$) correspond à la racine carrée du nombre de variables potentielles du modèle (noté $p$), soit $m = \sqrt{p}$. En d'autres mots, seulement un nombre limité de $m$ prédicteurs sont des candidats potentiels par nœud. Cela permet, s'il y a un prédicteur très fort dans l'échantillon, de créer des arbres avec d'autres variables comme premier nœud. Autrement dit, les arbres estimés de chaque échantillon bootstrap sont moins corrélés que lorsque toutes les variables sont prises en considération et donc la variance est réduite.

Comme énoncé plus haut, la méthode procède par *bootstrap*. Chaque arbre est donc estimé avec un échantillon différent. Il est donc possible, avec les observations inutilisées dans un échantillon *bootstrap*, de créer un échantillon de validation pour chaque arbre. On appellera ces erreurs de classification erreurs OOB (*out-of-bag*). Dans

ce travail, les erreurs OOB seront utilisées pour mesurer la performance de la procédure.

#### 2.2.3.3 Méthode *boosting*

Le *boosting* est une application simple d'une règle de prédiction en imposant de l'apprentissage à l'ordinateur. D'abord, un arbre avec trois nœuds est construit. Les erreurs de prédictions sont estimées pour toutes les observations. Un second arbre est construit avec trois nœuds, mais les observations avec des erreurs de prédictions importantes suite à la construction du premier arbre ont maintenant plus de poids dans le nouvel arbre que ceux qui n'ont pas eu d'erreurs. Le deuxième arbre aura possiblement d'autres variables dans la détermination de l'arbre que ceux du premier arbre. Un troisième arbre est alors construit avec de nouveaux poids et suivant la même technique. De nouvelles variables peuvent se montrer très utiles pour prédire des observations particulières. Il est important de choisir le nombre d'itérations approprié, qui sera déterminé par une minimisation des erreurs de classification par rapport au nombre d'itérations du modèle. Avec un nombre d'itérations assez large, la moyenne des prédictions peut être faite et on prend la classe la plus souvent prédite pour chaque observation *i*.

### 2.2.4 Calculs

Les estimations ont été produites avec la version 3.3.1 du logiciel libre *R*. Les paramètres des modèles utilisés sont ceux par défaut, indiqués dans la section des résultats ou en annexe. La liste ci-dessous présente les principales fonctions et *packages* utilisés pour procéder aux résultats obtenus dans cette recherche.

| Package | Description |
|---|---|
| Glmnet | *Lasso* en validation croisée, obtention du lambda optimal. |
| randomForest | Forêts aléatoires |
| Gbm | *Boosting* |
| ROCR | Différentes mesures de performance des modèles |

# CHAPITRE III

# DONNÉES

Les principales données proviennent de l'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ). Les variables environnementales proviennent de la base de données pancanadienne sur la qualité de l'air (BDPQA) et l'appariement se fera avec les fichiers de conversion des codes postaux (FCCP).

## 3.1 Données sur les enfants

L'ELNEJ est une enquête biennale menée par Statistique Canada sur les enfants canadiens suivant leur développement et leur bien-être de la naissance à la fin de l'adolescence. Les sujets englobent le comportement de l'enfant, l'éducation, la formation, l'apprentissage et la santé. L'information est recueillie par un intervieweur via des questionnaires et des tests cognitifs, auprès de la personne ayant la meilleure connaissance de l'enfant (ci-après, PCM). La collecte comporte huit cycles à intervalles de deux ans sur 35 795 enfants (au cycle huit) âgés de zéro à sept ans et de quatorze à vingt-cinq ans de 1994 à 2008.

Les enfants qui composent l'échantillon de la recherche sont ceux âgés de moins de deux ans au premier cycle (1994-1995) et qui se retrouvent dans trois cycles minimalement. Cet échantillon a été choisi, car il comporte des informations importantes sur la grossesse. Ces observations présentent des variables riches et précises sur les caractéristiques de naissance et de gestation de l'enfant.

### 3.1.1 L'état de santé de l'enfant observé à l'enquête

Les données comportent 9227 observations. La cohorte de départ au premier cycle (1994) recueille 2344 enfants de 0 ou 1 an. Au cycle 5, il y reste 1352 jeunes de 9 à 10 ans. Le tableau 3.1 montre les statistiques descriptives pour la variable de santé de l'enfant par rapport aux différents cycles utilisés.

TABLEAU 3.1 : Statistiques descriptives de la variable mauvaise santé

| | Cycle 1 | Cycle 2 | Cycle 3 | Cycle 4 | Cycle 5 |
|---|---|---|---|---|---|
| Année | 1994 | 1996 | 1998 | 2000 | 2002 |
| Âge | 0-2 ans | 2-4 ans | 5-6 ans | 7-8 ans | 9-10 ans |
| Observations | 2344 | 2141 | 1855 | 1535 | 1352 |
| Moyenne | 0,11 | 0,12 | 0,13 | 0,14 | 0,13 |
| Écart type | 0,32 | 0,33 | 0,34 | 0,34 | 0,33 |

La santé de l'enfant est déterminée sur une échelle de 1 à 5 (santé excellente, très bonne, bonne, passable ou faible). La variable est dichotomisée, prenant la valeur 1 si la santé de l'enfant est bonne, passable ou faible, 0 si très bonne ou excellente. Il est possible d'utiliser des modèles Probit ordonnées en gardant les cinq catégories de réponse, cela peut améliorer les prédictions en ajoutant de l'information. La décision a été prise en se basant sur les études antérieures et suite à la faiblesse des résultats préliminaires. La proportion des enfants en mauvaise santé a une tendance à la hausse, mais plutôt stable à partir du cycle 3. La figure 3.1 montre la proportion des enfants en mauvaise santé du cycle 2 à 5, selon leur état de santé à la naissance.

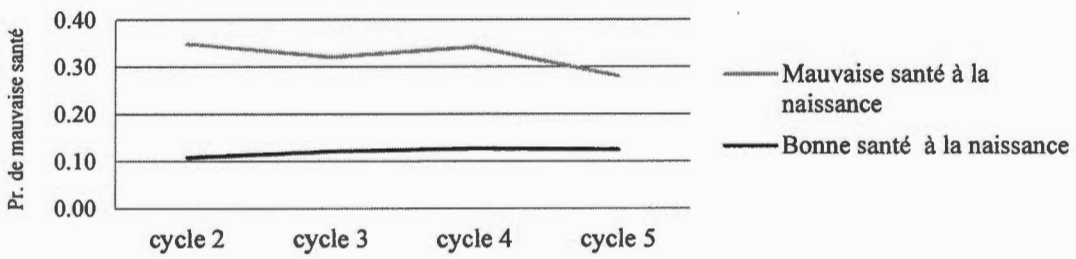



FIGURE 3.1 : Santé de l'enfant par rapport à la santé à la naissance

À la lumière de la figure 3.1, une autocorrélation de mauvaise santé est notable. En accord avec les études antérieures, il est plus probable pour un enfant en mauvaise santé à la naissance d'être en mauvaise santé par la suite. La variable de mauvaise santé au premier cycle représente la santé de l'enfant à la naissance, et non au moment de l'enquête. Cette variable dichotomique provient d'une réponse similaire à celle de la santé en général et subit la même transformation.

### 3.1.2 Statut socioéconomique

Plusieurs variables socioéconomiques sont retenues de l'ELNEJ. Le tableau 3.2 montre les moyennes de ces variables. Dans l'échantillon, la personne qui connait le mieux l'enfant (PCM et répondant du sondage) est la mère biologique dans 99,5 pour cent des cas. Les enfants étudiés sont des filles dans la moitié des cas. La variable dichotomique d'asthme indique si l'enfant a eu une crise d'asthme depuis sa naissance.

TABLEAU 3.2 : Statistiques descriptives des variables socioéconomiques de l'ELNEJ

| | Moyenne | Écart type |
|---|---|---|
| sexe (1 = fille) | 0,49 | 0,50 |
| Asthme | 0,05 | 0,22 |
| **Caractéristiques socioéconomiques** | | |
| Revenu | -0,09 | 1,01 |
| Éducation PCM (mère) | 0,65 | 0,48 |
| Éducation conjoint | 0,55 | 0,50 |
| **Caractéristiques familiales** | | |
| Mère est mariée | 0,71 | 0,46 |
| Deux parents bio | 0,88 | 0,33 |
| Mère fume | 0,32 | 0,47 |
| Mère immigrante | 0,09 | 0,28 |
| Conjoint est absent | 0,12 | 0,32 |
| Âge conjoint | 31,99 | 5,03 |
| Conjoint immigrant | 0,09 | 0,29 |
| Fratrie | 0,83 | 0,94 |
| Dépression parentale | 0,13 | 0,34 |
| Taille de la famille (log) | 1,31 | 0,26 |
| Maison unifamiliale Détachée | 0,33 | 0,47 |
| **Caractéristiques géographiques** | | |
| Densité population | 2,01 | 1,17 |
| Maritimes | 0,17 | 0,37 |
| Québec | 0,20 | 0,40 |
| Ontario | 0,26 | 0,44 |
| Prairies (SA / MA) | 0,16 | 0,37 |
| Alberta | 0,09 | 0,28 |
| Colombie-Britannique | 0,08 | 0,27 |

N=2344

La variable de revenu familial est exprimée en dollars réels canadiens (IPC de valeur 100 à l'an 2000), transformée en logarithme, centrée et réduite. Les valeurs manquantes

ont été imputées par Statistiques Canada par une technique de *bootstrap*. Une variable dichotomique d'imputation est ajoutée au modèle (15 % d'imputation, non montrée dans le tableau). La figure 3.2 montre les moyennes de la variable de mauvaise santé par cycle conditionnellement au tercile de revenu de la famille de l'enfant. Le premier tiers de revenu indique les revenus les plus faibles, le troisième les revenus plus élevés.

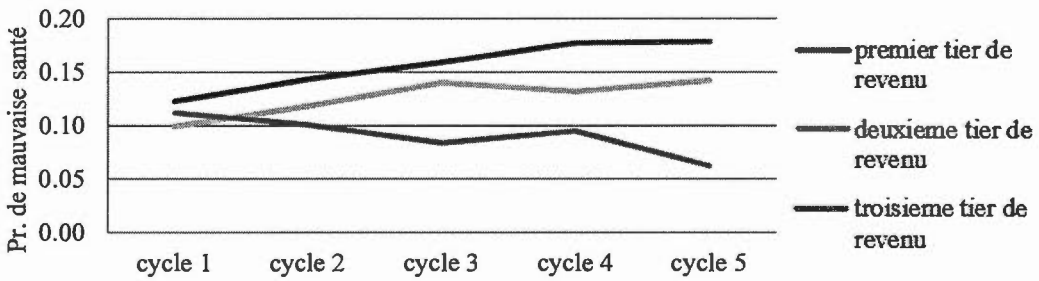



FIGURE 3.2 : Santé de l'enfant par rapport au revenu familial

Le revenu n'affecte pas la probabilité de mauvaise santé de l'enfant à la naissance. Pourtant, un écart se crée avec le temps. Au cycle 5, le fait d'appartenir au groupe de revenus faibles est associé à une probabilité d'environ 10 % de plus d'être en mauvaise santé en comparaison du groupe des plus riches. Les variables d'éducation, soient l'éducation de la mère et du conjoint, sont des variables dichotomiques qui prennent la valeur de 1 si le parent a fait des études postsecondaires (> D.É.S). La figure 3.3 montre la relation entre l'éducation de la mère et la probabilité du développement d'un problème de santé chez l'enfant.

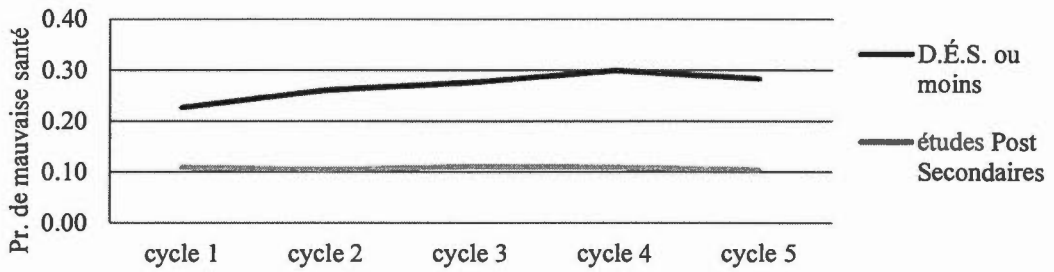



FIGURE 3.3 : Santé de l'enfant par rapport à l'éducation de la mère

La figure 3.3 permet de constater un écart important de l'espérance de mauvaise santé par rapport à l'éducation de la mère. Cela est en accord avec les études décrites plus haut et les résultats restent assez stables avec le temps.

Les variables familiales incluent l'état matrimonial (1 si mariée), d'immigration et de tabagisme pour la mère. Plusieurs variables pour le conjoint sont incluses, dont une d'absence du foyer. S'il n'y a pas de conjoint, toutes les variables dichotomiques le concernant sont mises à zéro. On considère aussi des variables pour la taille de la famille (log), le type de maison (1 si l'enfant vit dans une maison unifamiliale détachée), du fait d'avoir deux parents biologiques à la maison et de la fratrie. La figure 3.4 présente la probabilité d'être malade par rapport à l'état matrimonial de la mère.

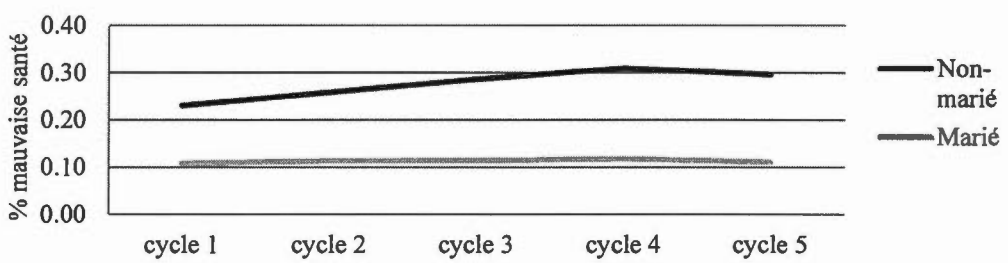



FIGURE 3.4 : Santé de l'enfant par rapport au statut marital de la mère

L'état matrimonial de la mère au premier cycle joue un rôle important dans la probabilité de mauvaise santé d'un enfant. Pour ce qui est de la dépression parentale, on retrouve dans l'ELNEJ le résultat d'un test de dépression familiale sur une échelle

de 36. Si le répondant ne voulait pas répondre au test, la note de 0 lui a été attribuée, mais une dichotomique de non-réponse a été ajoutée au modèle (15 % de non-réponse, la variable n'est pas incluse dans le tableau de statistiques descriptives). La note maximale étant de grands signes de dépression. Cette variable a été dichotomisée, prenant la valeur de 1 si le score du test est supérieur à 10, cela représente environ dix pour cent de la population. La figure 3.5 montre la relation entre la dichotomique de dépression et la santé de l'enfant.

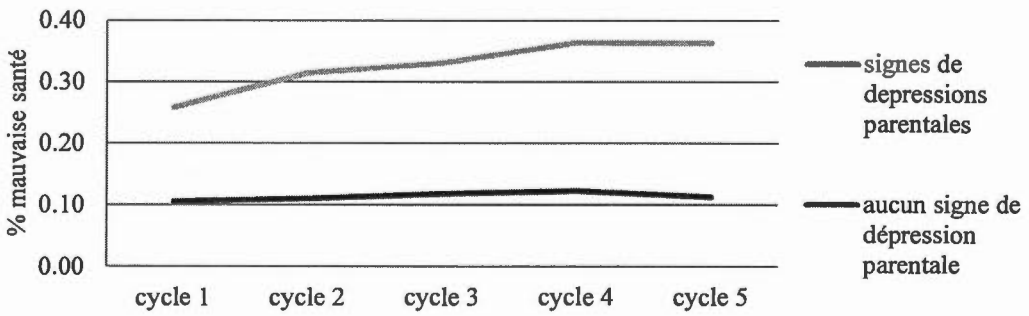



FIGURE 3.5 : Santé de l'enfant par rapport à la dépression parentale

Il est clair que les signes de dépressions chez les parents sont corrélés avec la probabilité, chez l'enfant, de développer des problèmes de santé. Cet écart semble s'agrandir avec le temps.

Pour les variables géographiques, on utilise une variable dichotomique avec quatre valeurs pour indiquer la densité de population dans lequel l'enfant réside, soit la ruralité, une région économique de moins de 100 000 habitants, une région économique entre 100 000 et 499 000 habitants, ou une région économique de plus 500 000

habitants[6]. Des dichotomiques de provinces (avec des regroupements pour les prairies et les maritimes) ont été ajoutées au modèle.

### 3.1.3 Caractéristiques de grossesse

Étant donné que les caractéristiques de la grossesse sont d'importants prédicteurs de la santé de l'enfant, l'échantillon ne comporte que les enfants avec de l'information à la grossesse. Le tableau 3.3 présente les statistiques descriptives relatives à la naissance de l'enfant.

---

[6] « Les provinces sont divisées en régions économiques (RÉ) [...]. Les régions économiques sont des régions géographiques de structure économique plus ou moins homogène constituées en vertu d'ententes fédérales-provinciales et qui sont relativement stables au fil du temps. » (*Enquête longitudinale nationale sur les enfants et les jeunes, cycle 8 — Guide de l'utilisateur*. P. 19. Section 5.2.2).

TABLEAU 3.3 : Statistiques descriptives relatives à la naissance de l'enfant

| | Moyenne | Écart type |
|---|---|---|
| **Durant la grossesse** | | |
| Drogues prescrites | 0,28 | 0,45 |
| Alcool | 0,17 | 0,38 |
| Tabac | 0,25 | 0,43 |
| Problèmes prénataux | 0,32 | 0,47 |
| **À la naissance** | | |
| Année de naissance | 3,50 | 0,55 |
| Mois de naissance | 6,43 | 3,31 |
| Âge de la mère | 28,10 | 4,93 |
| Mauvaise santé de la mère | 0,22 | 0,41 |
| Nombre de grossesses de la mère | 2,32 | 1,42 |
| Poids <2500 grammes | 0,05 | 0,22 |
| Naissance prématurée | 0,09 | 0,29 |

N=2433

L'année de naissance de l'enfant prend les valeurs de deux si l'enfant est né en 1992 à 5 si l'enfant est né en 1995. Les variables de grossesse indiquent si la mère a utilisé des drogues pharmaceutiques prescrites, de l'alcool (plus d'une fois par mois), le tabac ou si la mère a connu des problèmes prénataux. Les problèmes prénataux varient entre le diabète, l'hypertension et d'autres problèmes physiques nécessitant une intervention du médecin. La santé de la mère à la naissance est enregistrée, de la même manière que la santé de l'enfant. La transformation en variable dichotomique pour la santé se fait ensuite de la même façon. La figure 3.6 souligne la relation entre une mauvaise santé de la mère à la naissance et les problèmes de santé de l'enfant.

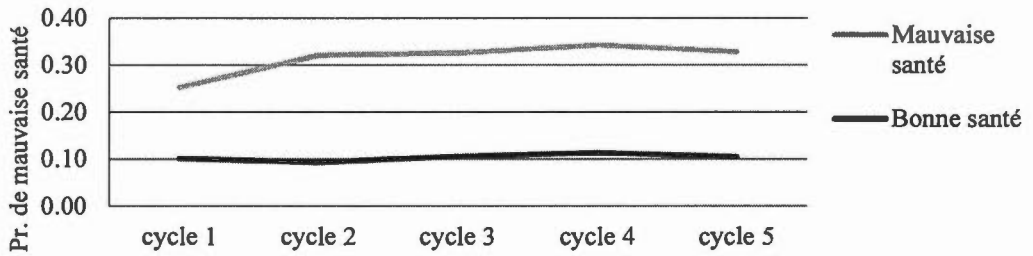



FIGURE 3.6 : Santé de l'enfant par rapport à la santé de la mère à la naissance

Le poids de l'enfant à la naissance (la normale est au minimum de 2500 grammes pour 94,3 % des enfants, modérément faible de 1500 à 2499 grammes pour 4,9 % des enfants ou très faibles à 1499 grammes ou moins pour 0,8 % des enfants) est transformé en variable dichotomique prenant la valeur de 1 si le poids de l'enfant est faible ou très faible. La variable de prématurité prend la valeur de 1 si l'enfant est né en 258 jours ou moins. La figure 3.7 montre le pourcentage des enfants en mauvaise santé par rapport à la durée de gestation de la grossesse de la mère.

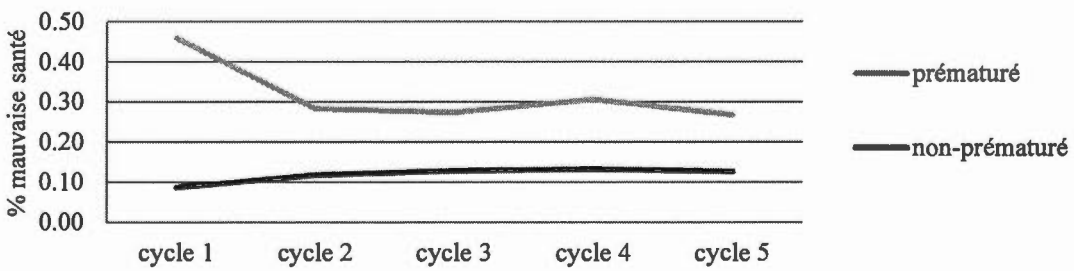



FIGURE 3.7 : Santé de l'enfant par rapport à la durée de gestation

On constate que les bébés prématurés à la naissance ont tendance à avoir une santé précaire en début de vie, mais l'écart se rétrécit avec le temps.

## 3.2 Données environnementales

La base de données pancanadienne sur la qualité de l'air (BDPQA) permet de capter les données collectées par le réseau national de surveillance de la pollution atmosphérique (RNSPA). Le programme RNSPA a été mis sur pied en 1969 et est géré au niveau provincial, mais normalisé au niveau national, afin de pouvoir fournir des données précises de la qualité de l'air dans les régions peuplées du Canada. Aujourd'hui, le réseau comprend 368 stations parmi 255 communautés, dans toutes les provinces et les territoires. Les données sont disponibles librement et le fichier des moyennes annuelles de 1994 a été utilisé comme données environnementales au premier cycle de l'ELNEJ.

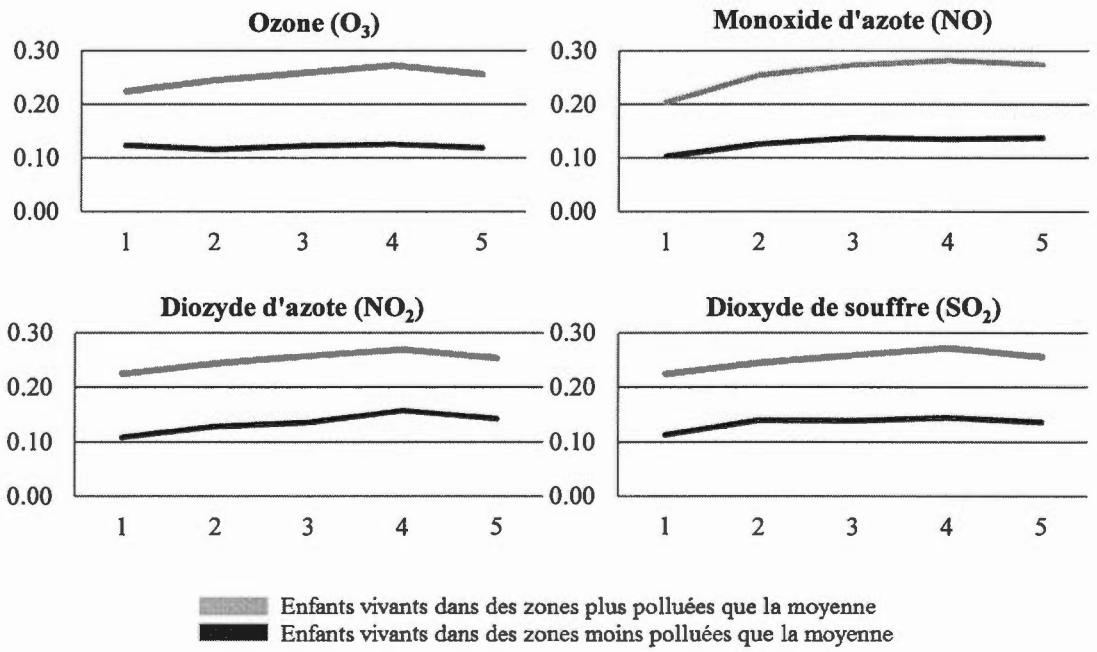
Bien que la BQPQA enregistre une multitude de polluants atmosphériques différents, les polluants choisis pour construire l'échantillon ont moins de 60 % de valeurs manquantes avant l'appariement. Les mesures utilisées, présentées au tableau 3.4, comprennent les différents niveaux d'ozone ($O_3$), trois mesures d'oxyde d'azote ($NO_x$, NO et $NO_2$), ainsi que le dioxyde de soufre ($SO_2$).

TABLEAU 3.4 : Statistiques descriptives des polluants et de l'imputation

| | $SO_2$ | $NO_2$ | $O_3$ | NO | $NO_X$ |
|---|---|---|---|---|---|
| % de captation | 93,01 | 82,82 | 89,60 | 82,82 | 82,94 |
| | (7,73) | (19,91) | (12,50) | (19,65) | (19,62) |
| Moyenne ($\mu g/m^3$) | 4,01 | 13,51 | 21,93 | 10,19 | 23,72 |
| | (3,10) | (6,67) | (5,37) | (13,57) | (19,55) |
| Écart type | 2,20 | 3,61 | 6,11 | 5,54 | 8,07 |
| | (1,41) | (1,23) | (1,54) | (7,10) | (7,66) |
| **Imputation** | | | | | |
| Disponibilités | 63 | 80 | 142 | 80 | 64 |
| Valeurs imputées | 96 | 79 | 17 | 79 | 95 |
| Bornes de captation | 159 | 159 | 159 | 159 | 159 |

Les écarts types sont entre parenthèses

Le pourcentage de captation exprime en pourcentage les jours de captation du polluant durant l'année 1994. La variable de moyenne est exprimée en microgramme par mètres cubes et représente la moyenne de pollution pour toutes les bornes pendant toute l'année, tandis que l'écart type représente les variations par borne. Au total, 159 bornes à travers le Canada captent les cinq différents polluants pour l'année 1994. L'imputation est discutée à la section 3.2.1. À la figure 3.8, on présente la relation entre quatre des cinq polluants et la variable de mauvaise santé de l'enfant. Sur chaque graphique, la ligne pâle représente les enfants vivants dans un environnement en haut de la pollution moyenne, tandis que la ligne plus foncée représente les enfants vivant dans un environnement moins pollué que la moyenne. Les axes horizontaux représentent les cycles de l'enquête, tandis que les axes verticaux représentent la probabilité de mauvaise santé.




Notes : Axe des ordonnées : Probabilité que l'enfant soit en mauvaise santé

FIGURE 3.8 : Santé de l'enfant par rapport aux différents niveaux de polluants

L'ozone ($O_3$) est un gaz présent à l'état naturel dans l'atmosphère (10 à 100 $\mu g/m^3$), par contre, avec une transformation dans l'atmosphère de certains polluants, comme les oxydes d'azote sous l'effet du rayonnement solaire, provoque des gênes respiratoires importantes.

L'oxyde d'azote ($NO_x$) se forme lors de phénomènes naturels, comme des orages ou des incendies, mais la principale source d'émanation est anthropique, dans la combustion des combustibles fossiles (charbon ou gaz naturel) ou les échappements d'automobiles (en particulier le diésel). Le monoxyde d'azote (NO) provient de la combustion à haute température d'azote. Ce gaz réagit avec de l'oxygène pour former du dioxyde d'azote ($NO_2$), essentiellement issu des sources de combustions automobile, industrielle et thermique (chauffage au gaz naturel, tabac). Le $NO_2$ est un gaz très toxique par inhalation.

Les dioxydes de soufre ($SO_2$) sont principalement issus d'activités humaines. L'inhalation est fortement irritante, il est libéré par de nombreux procédés industriels, comme la combustion de charbons, de pétrole et de gaz naturel, de même que dans la production de désinfectants, d'antiseptiques ou de conservation alimentaire. Une forte concentration de dioxyde de soufre provoque l'inflammation du système respiratoire. À la lumière de la figure 3.9, on constate que les quatre polluants présentés sont fortement liés avec les problèmes de santé de l'enfant.

Le fichier d'Environnement Canada contient aussi de l'information pertinente sur l'environnement météorologique qui peut aider à capter la pollution. Le tableau 3.5 tire de la BQPDA les autres variables utilisées dans la recherche.

TABLEAU 3.5 : Statistiques descriptives des autres variables de la BDQPA

| | Moyenne | Écart type |
|---|---|---|
| Latitude | 0,83 | 0,05 |
| Longitude | -1,47 | 0,34 |
| Élévation | 256,42 | 259,27 |
| Agricole | 0,06 | 0,24 |
| Commercial | 0,36 | 0,48 |
| Forestier | 0,07 | 0,25 |
| Industriel | 0,04 | 0,19 |
| Résidentiel | 0,36 | 0,48 |
| Vancouver | 0,01 | 0,12 |
| Montréal | 0,02 | 0,14 |
| Hamilton | 0,01 | 0,11 |
| Sudbury | 0,01 | 0,11 |
| Edmonton | 0,03 | 0,16 |
| Calgary | 0,04 | 0,20 |
| Distance | 68,04 | 105,93 |

N=157

L'altitude (Élévation) est une variable continue exprimée en mètres. Plusieurs variables dichotomiques de villes sont ajoutées, certaines d'entre elles (Toronto, Ottawa, etc.) ont été enlevées par souci de colinéarité et de convergence des modèles Logit et Probit. Une variable de type dichotomique de secteur est ajoutée; on voit une prévalence des enfants habitants proches de zones commerciales et résidentielles (72 % de l'échantillon). La variable de distance exprime en kilomètres la distance entre l'enfant et une borne environnementale. L'assignation des conditions environnementales se fait par une minimisation de la distance des bornes par rapport à chaque enfant et est discutée à la section 3.3. La distance moyenne entre les enfants et la borne environnementale la plus proche est de 68 km, avec un écart type très élevé.

### 3.2.1 Imputation des données

La commande `mi impute (regress)` du logiciel Stata 14.2 permet d'imputer les valeurs des variables continues manquantes avec une méthode d'imputation en chaîne markovienne. L'imputation permet une valeur renseignée pour chaque borne qui enregistre les polluants utilisés dans l'enquête.

La commande utilise une régression linéaire pour estimer les valeurs manquantes, la méthode markovienne exige une distribution normale des paramètres. L'imputation par régression linéaire est la méthode la plus populaire pour imputer des variables quantitatives continues. Les régresseurs utilisés pour l'imputation en chaîne sont les variables présentées au tableau 3.5. Pour vérifier si l'imputation donne des résultats plausibles, une différence de moyenne a été faite après l'imputation.

Les polluants imputés sont présentés au tableau 3.4. 366 observations ont été imputées au total, soit 46 % des observations totales. Le critère pour l'imputation est que chaque polluant devait avoir moins de 40 % de valeurs manquantes. Le critère d'imputation est justifié dans la mesure où l'imputation devient suspecte lorsqu'il y a plus de 50 % de valeurs manquantes. C'est alors que l'on observe un biais dans la moyenne et les écarts types des variables imputées (Scheffer, 2002).

## 3.3 Discussion de l'appariement

La BDPQA offre la position géographique de captation météorologique, tandis qu'on ne retrouve qu'une variable de code postal pour les enfants de l'ELNEJ. La transformation des codes postaux en données géographiques s'est fait avec le Fichier de Conversion des codes postaux (FCCP) de Statistique Canada. Ce fichier comprend plus de 800 000 codes postaux liés aux coordonnées de latitude et longitude, ainsi qu'un

code de transformation qui est inclus et utilisé avec le logiciel SAS (version 7). La formule de transformation des codes postaux en données géographiques euclidiennes utilisée est la formule standard Haversine. La distance entre chaque enfant et sa borne géographique est mesurée et la borne géographique choisie est celle avec la distance la plus proche.

Les bases de données ont été importées, appariées et nettoyées avec le logiciel *Stata* version 14. Les codes postaux des enfants ont été transformés en données géographiques avec le code fourni par Poste Canada, en utilisant la version SAS *version* 7. La minimisation de distance géographique, ainsi que les résultats ont été obtenus avec la version 3.3.1 du logiciel libre *R*.

# CHAPITRE IV

# RÉSULTATS

Les résultats de la recherche sont présentés de la manière suivante; tout d'abord, la performance des modèles est comparée avec et sans variables environnementales, ensuite, les modèles les plus performants sont retenus pour explorer le pouvoir prédictif des différentes variables utilisées.

## 4.1 Résultats des différentes méthodes

Le tableau 4.1 présente les différents résultats des méthodes présentées dans la recherche. Quatre mesures de performances sont présentées, soit l'erreur de classification, la spécificité, la sensibilité et l'aire sous la courbe (*AUC*). Pour chaque modèle, la colonne de gauche indique la performance sans variables environnementales, soit avec 39 variables ou moins de l'ELNEJ, tandis que les colonnes de droites incluent les variables environnementales (un total de 73 variables ou moins). Les prédictions sont faites à partir des données du premier cycle au cycle deux (2141 observations), au cycle trois (1855 observations), au cycle quatre (1535 observations) et au cycle cinq (1352 observations). Pour les erreurs de classification, la spécificité et la sensibilité, le critère bayésien est utilisé pour classifier les prédictions, soit une prédiction positive de mauvaise santé lorsque la probabilité prédite est supérieure à 0,5. Des colonnes de moyennes ont aussi été ajoutées au modèle par cycle, ainsi que par modèles.

Pour l'ajustement du modèle lasso, le λ optimal est trouvé par validation croisée ($k$ = 10) et le critère de performance choisit est le AUC, les résultats graphiques sont présentés en annexe. Les λ optimaux varient entre 0,0010 et 0,0038 (tableau 4.2). Pour l'ajustement du boosting, le nombre d'itérations varie entre 730 et 1227, la méthode nécessite moins d'itérations lorsque l'horizon de prédiction est éloigné.

TABLEAU 4.1 : Performance des modèles

| | Logit | | Probit | | Lasso | | Forêt aléatoire | | Boosting | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Non | Oui | Non | Oui | Non | Oui | Non | Oui | Non | Oui | Moy. |
| **Erreur de classification** | | | | | | | | | | | |
| Cycle 2 | 0,85 | 0,83 | 0,85 | 0,83 | 0,86 | 0,89 | 0,85 | 0,85 | 0,89 | 0,89 | 0,86 |
| Cycle 3 | 0,85 | 0,82 | 0,85 | 0,82 | 0,88 | 0,87 | 0,85 | 0,85 | 0,88 | 0,88 | 0,85 |
| Cycle 4 | 0,85 | 0,82 | 0,85 | 0,82 | 0,87 | 0,87 | 0,85 | 0,81 | 0,87 | 0,88 | 0,85 |
| Cycle 5 | 0,83 | 0,80 | 0,82 | 0,80 | 0,86 | 0,86 | 0,85 | 0,81 | 0,86 | 0,86 | 0,83 |
| Moyenne | 0,84 | 0,82 | 0,84 | 0,82 | 0,87 | 0,87 | 0,85 | 0,83 | 0,87 | 0,88 | 0,85 |
| **Sensibilité** | | | | | | | | | | | |
| Cycle 2 | 0,01 | 0,04 | 0,01 | 0,04 | 0,01 | 0 | 0 | 0,03 | 0 | 0 | 0,01 |
| Cycle 3 | 0,07 | 0,09 | 0,07 | 0,09 | 0 | 0 | 0 | 0 | 0 | 0 | 0,03 |
| Cycle 4 | 0,08 | 0,08 | 0,08 | 0,08 | 0 | 0 | 0 | 0 | 0 | 0 | 0,03 |
| Cycle 5 | 0,02 | 0,03 | 0,02 | 0 | 0 | 0 | 0 | 0,02 | 0 | 0 | 0,01 |
| Moyenne | 0,04 | 0,06 | 0,04 | 0,05 | 0,01 | 0 | 0 | 0,01 | 0 | 0 | 0,02 |
| **Spécificité** | | | | | | | | | | | |
| Cycle 2 | 0,99 | 0,98 | 0,99 | 0,98 | 0,99 | 1 | 1 | 0,99 | 1 | 1 | 0,99 |
| Cycle 3 | 0,99 | 0,98 | 0,99 | 0,98 | 1 | 1 | 1 | 1 | 1 | 1 | 0,99 |
| Cycle 4 | 0,99 | 0,98 | 0,99 | 0,98 | 1 | 1 | 1 | 1 | 1 | 1 | 1,00 |
| Cycle 5 | 0,99 | 0,98 | 1 | 0,98 | 1 | 1 | 1 | 0,99 | 1 | 1 | 1,00 |
| Moyenne | 0,99 | 0,98 | 0,99 | 0,98 | 0,99 | 1 | 1 | 0,99 | 1 | 1 | 0,99 |
| *AUC* | | | | | | | | | | | |
| Cycle 2 | 0,68 | 0,68 | 0,69 | 0,69 | 0,70 | 0,69 | 0,66 | 0,65 | 0,71 | 0,71 | 0,69 |
| Cycle 3 | 0,64 | 0,63 | 0,64 | 0,63 | 0,62 | 0,61 | 0,65 | 0,63 | 0,63 | 0,61 | 0,63 |
| Cycle 4 | 0,65 | 0,65 | 0,66 | 0,66 | 0,65 | 0,65 | 0,62 | 0,60 | 0,64 | 0,63 | 0,64 |
| Cycle 5 | 0,55 | 0,50 | 0,55 | 0,51 | 0,65 | 0,65 | 0,57 | 0,54 | 0,64 | 0,61 | 0,58 |
| Moyenne | 0,63 | 0,62 | 0,64 | 0,62 | 0,66 | 0,65 | 0,63 | 0,61 | 0,66 | 0,64 | 0,63 |

Note : Pour chaque méthode, la colonne *non* représente les performances sans ajout de variables environnementales; la colonne *oui* représente les performances avec l'ajout de variables environnementales. La spécificité représente la proportion des enfants en bonne santé correctement identifiée par la méthode, la sensibilité représente la proportion des enfants en mauvaise santé correctement identifiés par la méthode.

Les modèles classifient correctement la santé de l'enfant à 85 % du temps en moyenne. En regardant uniquement les erreurs de classification, les modèles

performent de façon assez similaire. Le meilleur modèle est le *boosting* avec les variables environnementales qui classifie correctement la santé de l'enfant à 89 % pour la prédiction au cycle 2, et 88 % du temps pour les cycles subséquents. Pour ce modèle, l'ajout de variables environnementales améliore d'un point de pourcentage la classification de la santé de l'enfant en moyenne. Le deuxième meilleur modèle est le *lasso*, qui classifie correctement la santé de l'enfant à 87 % du temps, avec et sans variables environnementales. En moyenne, ce modèle améliore les prédictions de trois points de pourcentage par rapport au Logit et au Probit qui contiennent toutes les variables, et de cinq points de pourcentage lorsque l'on rajoute les variables environnementales. En suivant les erreurs de classification, les prédictions perdent trois pour cent de précision en changeant l'horizon de prévision du deuxième cycle au dernier. Les autres modèles performent moins bien lorsque les variables environnementales sont incluses.

Pour tous les modèles, la sensibilité est faible et la spécificité est haute. On observe même une sensibilité de 0 pour le *boosting*, le *lasso* avec les variables environnementales et la forêt aléatoire sans variables environnementales. Cela signifie que le modèle classe tous les enfants en bonne santé ($\hat{y}_i = 0 \; \forall \; i$ ). Dans une perspective où on souhaite pénaliser pour les erreurs fausses négatives, il est important d'utiliser la mesure $AUC$ comme critère de détermination des meilleurs modèles.

Suivant le critère $AUC$, les modèles performent similairement sur un horizon rapproché, soit au cycle deux. C'est lorsque l'on regarde au cycle cinq que la performance des modèles se distingue. Les deux meilleurs modèles sont les modèles *boosting* et *lasso* sans variables environnementales avec un $AUC$ de 0,66. En moyenne, les modèles sont plus performants sur un horizon temporel plus court, l'aire sous la courbe passe, en moyenne, de 0,69 à 0,58 lors du changement de prévision du cycle 2 au cycle 5.

Au cycle 5, les régressions du lasso performent mieux par rapport au Logit et au Probit de 10 points sans variables environnementales et de 15 points avec.

Pour tous les modèles, l'ajout de variables environnementales n'améliore pas la mesure *AUC*. Les modèles *boosting* et *lasso* sont choisis pour déterminer les variables les plus puissantes pour la prédiction. Ainsi, il sera possible de déterminer la force de prédictions des variables.

## 4.2 Variables retenues des méthodes choisies

Les variables choisies par le modèle *lasso* sont celles qui sont les meilleures pour prédire, en validation croisée, la variable dépendante du modèle. Le choix du lambda est intrinsèquement relié au nombre de variables optimales. Le lambda optimal a été choisi en validation croisée, les résultats sont présentés en annexe (figures A.1 et A.2). Le tableau 4.2 présente les coefficients des régressions *lasso* avec et sans variables environnementales. Pour chaque cycle, on présente le nombre de variables retenues. Si une variable est retenue dans plus de deux cycles pour un modèle, les coefficients de la variable sont alors présentés.

Les écarts types ne sont pas présentés, car il est impossible de donner une signification réelle aux coefficients des régressions. La raison pour cela est que les méthodes d'estimations pénalisées, tel le *lasso*, sont des procédures qui réduisent la variance des estimateurs en introduisant des biais substantiels. Le biais de chaque estimateur est, donc, la partie principale de l'écart type des coefficients. En conséquence, les écarts types ne sont pas significatifs pour des coefficients fortement biaisés. Malheureusement, il est impossible d'obtenir un estimé précis du biais (Goeman, Meijer et Chaturvedi, 2016).

En conséquence, il est impossible de calculer la marginalité des coefficients. Le signe du coefficient reste quand même un indicateur de prédiction, ainsi que la présence de la variable dans le modèle.

TABLEAU 4.2 : Résultats de la méthode *lasso*

| | Sans variable environnementale | | | | Avec variables environnementales | | | |
|---|---|---|---|---|---|---|---|---|
| | Cycle 2 | Cycle 3 | Cycle 4 | Cycle 5 | Cycle 2 | Cycle 3 | Cycle 4 | Cycle 5 |
| $\Lambda$ | 0,001 | 0,016 | 0,019 | 0,038 | 0,037 | 0,013 | 0,021 | 0,031 |
| Nombre Variables | 34 | 11 | 8 | 3 | 4 | 13 | 7 | 4 |
| Constante | -3,822 | -2,129 | -1,866 | -2,016 | -2,03 | -2,353 | -1,853 | -2,02 |
| Dépression parentale | 0,308 | 0,221 | 0,219 | . | . | 0,277 | 0,195 | . |
| Santé naissance | 0,803 | 0,267 | 0,203 | . | 0,109 | 0,341 | 0,159 | . |
| Mère tabac | 0,076 | 0,204 | . | . | . | 0,238 | . | . |
| Habitation | 0,241 | . | 0,015 | . | . | . | . | . |
| Asthme | 1,526 | 0,907 | 0,946 | . | 0,805 | 0,965 | 0,905 | . |
| Revenu | 0,045 | -0,079 | -0,045 | -0,048 | . | -0,074 | -0,039 | -0,094 |
| Éducation conjoint | -0,143 | -0,094 | -0,301 | . | . | -0,12 | -0,28 | . |
| Santé Mère (naiss.) | 0,645 | 0,4 | 0,386 | 0,153 | 0,274 | 0,422 | 0,366 | 0,265 |
| Prairies (loc.) | -0,312 | -0,01 | . | . | . | . | . | . |
| N | 2141 | 1855 | 1535 | 1352 | 2141 | 1855 | 1535 | 1352 |

La variable de mauvaise santé de la mère, à la naissance de l'enfant, est la seule variable qui est retenue pour toutes les régressions. Une mauvaise santé de la mère, dans tous les cas, a tendance à augmenter la probabilité que l'enfant se classifie en mauvaise santé. La deuxième variable la plus importante est la variable de revenu parental, qui se qualifie dans sept des huit régressions. En général, une augmentation de revenu diminue la probabilité que l'enfant soit en mauvaise santé. Les variables qui se retrouvent six fois sont l'asthme au premier cycle et la mauvaise santé à la naissance de l'enfant. On observe aussi l'éducation du conjoint et le score de dépression parentale qui jouent un rôle important à cinq reprises.

En prenant le critère *AUC,* la méthode *boosting* est le meilleur modèle de prédiction de la recherche. Il est possible de capturer l'influence relative de chaque variable par la méthode de Friedman (2011). Cette mesure d'importance relative est mesurée avec la réduction de la variance en utilisant l'index Gini (le critère de sélection). La mesure est alors exprimée relativement à la variable la plus influente.

La mesure d'influence des variables n'est pas linéaire. Les premières variables sont très influentes par rapport aux autres. La figure 4.1 ne montre que les quinze variables les plus influentes, car l'influence des autres variables est négligeable. Le même tableau pour le *boosting* avec les variables environnementales est présenté en annexe (A.3). Le tableau 4.3 montre les variables par ordre d'influence relative pour la méthode *boosting.*

TABLEAU 4.3 : Variables principales de la méthode *boosting* par ordre d'influence relative sans variables environnementales

| Cycle 2 | | Cycle 3 | | Cycle 4 | | Cycle 5 | |
|---|---|---|---|---|---|---|---|
| Asthme | 16,5 | Revenu | 19,1 | Asthme | 23,7 | Revenu | 25,6 |
| Santé Naissance | 15,2 | Santé Mère | 16,7 | Revenu | 11,7 | Santé Mère | 14,2 |
| Santé Mère | 11,6 | Mois de naiss. | 9,9 | Santé Mère | 11,4 | Dép. parentale | 13,3 |
| Grosseur famille | 7,0 | Mère tabac | 8,1 | Mois de naiss. | 5,8 | Québec (loc) | 7,9 |
| Grossesses | 6,0 | Grossesses | 6,2 | Dépression parent | 5,3 | Âge mère naiss. | 4,6 |
| Mois de naiss. | 5,2 | Mère mariée | 5,5 | Éduc. conjoint | 5,0 | Mois de naiss. | 4,6 |
| Taille pop. | 4,7 | Âge mère naiss. | 4,9 | Mère mariée | 5,0 | Naiss. p. poids | 3,6 |
| Fratrie | 3,5 | Asthme | 4,0 | Santé Naissance | 4,0 | Drogues Presc. | 2,9 |
| Âge mère naiss. | 3,5 | Habitation | 3,5 | Éduc. mère | 3,3 | Grossesses | 2,8 |
| Revenu | 3,5 | Éduc. Conjoint | 1,9 | Âge mère naiss. | 2,9 | Éduc. mère | 2,7 |
| Mère cigarette | 3,2 | Grosseur famille | 1,5 | Fille | 2,8 | Fille | 2,0 |
| Dépression parent | 2,6 | Fratrie | 1,5 | Grosseur famille | 2,1 | Fratrie | 1,7 |
| Fille | 2,1 | C.-B. (loc.) | 1,5 | Grossesses | 2,0 | Prob. Prénat, | 1,6 |
| Habitation | 1,8 | Éduc. Mère | 1,4 | Revenu imp. | 1,9 | Taille pop. | 1,3 |
| Drogues Presc. | 1,4 | Drogues Presc. | 1,3 | Drogues Pharma. | 1,7 | Grosseur famille | 1,1 |

La variable la plus influente au cycle 2 et au cycle 4 est l'asthme (16,5 % et 23,7 % d'amélioration totale du modèle). Aux cycles 3 et 5, c'est le revenu parental qui est la variable la plus influente (19,1 % et 25,6 %). Aussi, la santé de la mère à la naissance figure à tous les cycles dans les trois variables les plus influentes. Parmi les autres variables, on retrouve dans tous les cycles la variable du nombre de grossesses de la mère, le logarithme de la taille de la famille, le mois de naissance de l'enfant, l'âge de la mère à la naissance. Graphiquement, les figures en annexe A.3 et A.4 montrent l'influence relative des modèles présentés. En haut à gauche présente l'influence relative des variables pour les prédictions au deuxième cycle, en haut à droite du troisième cycle, en bas à gauche au quatrième cycle et en bas à droit au dernier cycle. Sur ces figures, il est possible d'observer graphiquement la forme exponentielle de l'influence des variables. En fait, plus de 50 % de la variation de la santé de l'enfant peut être expliquée par les quatre premières variables.

Lors de leur ajout, les variables environnementales ne figurent jamais, pour toutes les régressions, dans au moins une des trois variables les plus influentes (le tableau similaire au tableau 4.3 est présenté en annexe). Aucune variable environnementale ne dépasse 5 % de contribution à l'amélioration du modèle. La mauvaise santé de la mère à la naissance reste dans les trois variables les plus influentes du modèle. Le logarithme du revenu et l'asthme de l'enfant restent dans toutes les régressions comme une des variables les plus influentes. Le logarithme de la taille de la famille, le mois de la naissance de l'enfant et l'âge à la naissance de la mère deviennent des variables négligeables. Ces variables sont remplacées par des variables environnementales, notamment les variables qui ont à trait à l'ozone.

Les résultats détaillés des autres méthodes sont présentés en annexe.

# CHAPITRE V

# DISCUSSION

## 5.1 Discussion sur la performance des modèles

Lors de la prédiction de la santé de l'enfant à l'âge de 4 ans, toutes les méthodes utilisées performent d'une manière assez similaire. C'est lors des prédictions de la santé à l'âge de 10 ans que l'on commence à observer des différences de performance entre les modèles. À partir des variables à la naissance, il est possible d'estimer correctement la santé de l'enfant à 85 % du temps en moyenne à l'âge de 10 ans. La prédiction de la santé de l'enfant à 4 ans n'est que 3 % plus précise que celle à l'âge de 10 ans.

De par les mesures de sensibilité et sensibilité des modèles, il est possible de constater que la performance en termes des erreurs de classification s'explique parce que les modèles ont tendance à classer en quasi-totalité les observations en bonne santé. Cela est naturel, puisque la majorité des enfants sont en bonne santé, mais dans une optique où il y a un arbitrage entre la sensibilité et la spécificité (discuté à la section 5.4), le critère de performance optimal ne peut pas être les erreurs de classification.

Un résultat intéressant de l'étude se trouve dans les résultats du modèle *lasso*. Il est possible de constater que les modèles optimaux contiennent un nombre restreint de variables (moins de quinze variables pour sept des huit modèles de prédiction). Surtout lors des prédictions des cycles plus avancées, les modèles avec moins de variables sont les plus performants.

Avec le critère *AUC,* les modèles qui se distinguent dans les prévisions à long terme sont le *lasso* et le *boosting*. Cela confirme qu'en termes de performance hors

échantillon, utiliser trop de variables dans des modèles de prédiction risque de suridentifier le modèle. Aussi, les résultats soutiennent que les méthodes d'apprentissage automatique, tel le *boosting*, performent mieux que les régressions en prédiction binomiale. Si la méthode de forêts aléatoires fait moins bien que le *lasso*, cela indique qu'il n'y a pas beaucoup d'effets d'interactions entre les variables.

La performance générale des modèles est difficile à classer. Nous savons qu'un AUC de 0,5 est une mauvaise prédiction, et qu'un AUC de 1 est une excellente. Par contre, il est difficile de savoir si la méthode est bonne, excellente ou passable, car cela dépend du contexte. Certains auteurs disent que la mesure AUC n'est qu'une mesure relative et, par conséquent, ne permet que de comparer des modèles entre eux. D'autres auteurs (Tape, T. G, 2017) proposent le système académique de points pour classifier la performance, soit ;

- 0,90 à 1 = Excellent (A)
- 0,80 à 0,90 = Bonne (B)
- 0,70 à 0,80 = Passable (C)
- 0,60 à 0,70 = Faible (D)
- 0,50 à 0,60 = Échec du modèle (F)

Suivant cette échelle, les performances des modèles sont faibles ou en échec. Étant donné que plusieurs modèles sont testés, il serait surprenant que ce soit la méthode qui fait défaut, plutôt qu'un problème dans les données. Les problèmes de données seront discutées aux sections 5.2 et 5.3. Une recommandation pour les prochaines études serait d'inclure des variables comme les dossiers médicaux et la génétique pour améliorer potentiellement les prédictions des modèles. Toutefois, la performance des modèles est similaire à celle des études de prédictions similaires (Chittleborough *et al.* ,2010).

## 5.2 Discussion sur les variables principales de l'étude

La santé de la mère est la seule variable qui se retrouve dans tous les cycles du *lasso*, de même que dans les trois variables les plus prédictives dans tous les modèles du *boosting*. Par conséquent, la santé de la mère à la naissance de l'enfant est la variable qui prédit le mieux la santé de l'enfant dans l'étude. Parmi les autres variables importantes des modèles retenus, on note le revenu familial, qui se retrouve dans sept des huit régressions *lasso* comme variables retenues et dans les trois variables les plus importantes dans six des huit régressions *boosting*, ainsi que la santé de l'enfant à la naissance et l'asthme durant le premier cycle. Ces résultats sont utiles pour confirmer les études causales dans un contexte de prédiction.

Pour la méthode *lasso* (tableau A3), les seuls prédicteurs qui sont retenues dans les prédictions pour tous les cycles sont la santé de la mère à la naissance et la variable de revenu. Les valeurs des coefficients restent assez stables à travers les prédictions. Pour la méthode *boosting*, la variable sur la santé de la mère reste parmi les trois variables les plus puissantes prédictives pour tous les cycles. La variable de revenue familial reste aussi importante dans tous les cycles. Les variables prédictives principales sur la santé de l'enfant sont, donc, la santé de la mère à la naissance, ainsi que le revenu familial, et ce, pour tous les cycles. Certaines variables environnementales apparaissent parmi les quinze meilleures variables prédictives pour un cycle, mais aucune ne reste dans le groupe des meilleures variables prédictives sur la période des cinq cycles.

Contrairement aux études causales, l'éducation parentale n'est pas retenue comme un prédicteur particulièrement important de la santé de l'enfant. Cela est peut-être dû au fait que l'éducation parentale est corrélée avec le salaire familial et que la variation de cette variable capte celle de l'éducation.

Suite aux résultats de la recherche, une recommandation politique possible est de faire passer un test de dépression parentale durant la grossesse. Effectivement, la dépression

parentale est un excellent prédicteur de la santé de l'enfant et est facile à capter, à l'aide d'un test. C'est une technique peu coûteuse et effective. Par exemple, un test de dépression pourrait être distribué durant la grossesse de l'enfant. Ainsi, il serait possible, avec les résultats, d'identifier rapidement les enfants à risque.

## 5.3 Discussion sur l'ajout des variables environnementales

L'ajout de variables environnementales entraîne la surspécification des modèles, car la performance *AUC* décline pour tous les modèles lors de l'ajout de variables environnementales. Ces résultats sont possiblement liés à plusieurs raisons;

a) L'effet de la pollution peut être capté par d'autres variables corrélées

Une des difficultés d'isoler les effets des variables sur la santé de l'enfant est la haute corrélation entre les variables indépendantes. L'éducation parentale, le revenu familial et l'endroit où les parents décident d'habiter sont fortement corrélés. Par exemple, les familles plus pauvres ont tendance à s'installer dans des quartiers industriels, donc plus pollués. Par conséquent, il se peut que la variable de revenu capte l'effet de la qualité du milieu. Une fois que des variables importantes sont incluses dans le modèle, comme celles du revenu de la santé à la naissance, la pollution n'ajoute pas d'information pertinente à la prédiction.

b) Mauvaise minimisation de distance

Le calcul de minimisation de la distance pourrait causer des difficultés dans l'étude, car elle ne prend pas en considération deux aspects importants. D'une part, aucune pénalité n'a été imposée par rapport aux enfants se trouvant loin des bornes

météorologiques, cela pourrait avoir comme conséquence que la pollution captée ne reflète pas l'exposition réelle de l'enfant. D'autre part, la minimisation de distance n'a pas pris en considération des barrières naturelles de pollution, tels des fleuves ou des montagnes.

c) Problème avec les données

Les données utilisées sont des données agrégées sur une base annuelle. Or, il se peut que les données ne soient pas assez précises pour bien capter l'effet de la pollution sur la santé de l'enfant. Il est possible que l'impact de la pollution soit relié à des extrêmes ou à des effets d'interaction de plusieurs polluants ou de polluants avec des températures élevées. Il est démontré que l'humidité joue un rôle dans la volatilité des polluants et, par conséquent, devrait être prise en considération lors des études futures. De plus, utiliser des données sur une base journalière permettrait de capter les variations de pollutions plus précisément.

## 5.4 Recommandations politiques

Bien connaître les facteurs de risques de la santé de l'enfant permet à un gouvernement de faire une analyse coûts-bénéfices dans le but d'instaurer des politiques de santé publique.

Un exemple de politique serait une aide médicale et préventive particulière aux enfants plus à risque de développer des problèmes de santé. Dans ce cas, le coût fixe du projet comprendrait la récolte d'information sur les femmes enceintes (par exemple, tel que mentionné ci-haut, un test de dépression parental pour tous), ainsi que le coût d'une

telle politique au niveau gouvernemental. Le coût variable de la politique serait ceux des médicaments, traitements et de salaires supplémentaires par des médecins ou des intervenants sociaux par enfant dépisté.

Créer une telle politique publique prendrait une étude causale distincte, car ce n'est pas suffisant d'identifier les enfants à risque, mais il faut que l'aide apportée soit efficace.

Le bénéfice de ce projet serait de prévenir le développement de mauvaise santé de plusieurs enfants. Il est difficile de savoir combien coûte un enfant malade à la société. En plus des frais médicaux, plusieurs coûts indirects sont rattachés à la mauvaise santé de l'enfant, comme le nombre de jours d'absence au travail des parents, l'impact sur la réussite scolaire de l'enfant, les coûts futurs d'une mauvaise santé dans le système, etc.

De plus, l'analyse dépend de la qualité des prédictions. La prévention qu'entraîne une bonne prédiction (un enfant malade dépisté) ne sera pas garante de succès. De plus, la prévention qu'entraîne une mauvaise prédiction (un enfant malade non dépisté ou un enfant en santé dépisté) entraînera une politique inefficace ou des coûts d'intervention inutiles. Savoir quels sont les coûts des erreurs fausses positives et fausses négatives est important pour doser l'arbitrage entre la spécificité et la sensibilité souhaitées. Si, par exemple, le coût d'intervention est faible, il serait préférable de choisir un modèle qui minimise les erreurs de spécification, dans le but de pénaliser les enfants en mauvaise santé, aux dépens de la sensibilité.

Le calcul de la valeur actuelle nette d'une telle politique est difficile à faire, car beaucoup d'éléments sont à considérer. Une des extensions possibles de ce mémoire pourrait être de quantifier les coûts directs et indirects d'une telle politique, ainsi que les bénéfices attachés à son imposition.

# CONCLUSION

L'objectif principal de ce mémoire est de mesurer l'apport de variables environnementales sur la santé de l'enfant. Pour se faire, la cohorte longitudinale de l'ELNEJ a été appariée à des bornes météorologiques par une minimisation de distance géographique pour capter la pollution dont l'enfant est exposé à sa naissance. Par rapport aux variables de la naissance, des prédictions à l'âge de 4, 6, 8 et 10 ans ont été faites en utilisant des régressions Logit et Probit et des modèles d'apprentissage automatique, tels *lasso,* forêts aléatoires et *boosting.* Les problèmes à résoudre ont été l'appariement des données, la minimisation des distances géographiques et le critère de sélection de performance des modèles.

C'est à un stade plus avancé de l'enfance que les prédictions des modèles commencent à se distinguer. Suivant le critère *AUC*, les deux modèles les plus performants sont le *lasso* et le modèle *boosting*. La variable la plus importante pour prédire la santé de l'enfant est la santé de la mère à la naissance de l'enfant. Les variables de revenus, et de santé de l'enfant à la naissance, tel l'asthme ou de santé générale, sont aussi d'excellents prédicteurs. Toutefois, il semble que les modèles performent pauvrement, il se pourrait qu'il y a un aspect de chance qui vienne jouer dans le développement de maladies chez l'enfant. Pour améliorer la performance des modèles, l'apport de variables sur la génétique de l'enfant serait bénéfique.

L'ajout de variables environnementales n'améliore pas les performances globales des modèles. Étant donné que la mesure *AUC* décline lors de l'ajout de variables environnementales, il est possible qu'elles causent une surspécification des modèles. Pour améliorer la qualité de l'apport des variables environnementales dans les prédictions, une série de modifications sur les données peut être faite, comme une fréquence d'enregistrement plus courte et une minimisation de distance tenant compte des différents changements.

Les recommandations politiques proposées dépendent des coûts associés à l'imposition de programme de prévention de la santé de l'enfant. Un tel programme imposerait un test de dépression parentale durant la grossesse, ainsi qu'une analyse des avantages et des coûts par rapport au statut de l'enfant.

En somme, les méthodes de prédictions sont encore souvent négligées en sciences économiques. Cependant, avec les bases de données grandissantes, l'accessibilité de la genèse humaine et le pouvoir de calcul des ordinateurs modernes, les méthodes d'apprentissage automatique sont des moyens pertinents de valider des intuitions économiques. Dans ce mémoire, l'application de ces méthodes a permis de valider l'intuition économique, ainsi que les hypothèses émises par les économistes dans les recherches causales. Toutefois, l'exploration de ces méthodes et le raffinement de ces données, proposées comme extension de recherche, sont pertinents.

# BIBLIOGRAPHIE


Allin, S. et Stabile, M. (2012). Socioeconomic Status and Child Health: What Is the Role of Health Care, Health Conditions, Injuries and Maternal Health? *Health Economics, Policy and Law* 7, 227-242.

Barnett, A., Williams, G.M., Schwartz, J., Neller, A.H., Best, T.L., Petroeschevsky, A.L. et Simpson, R.W. (2005). Air Pollution and Child Respiratory Health: A Case-Crossover Study in Australia and New Zealand. *American Journal of Respiratory and Critical care medicine* 51, 1272-1278.

Case, A., Fertig, A. et Paxson. C. (2005). The Lasting Impact of Childhood Health and Circumstances. *Journal of Health Economics* 24(2) 365-389.

Case, A., Lubotsky, D. et Paxson. C. (2002). Economic Status and Health in Childhood: The Origins of the Gradient. *American Economic Review* 92(5), 1308–34.

Chittleborough, C.R., Searle, A.K., Smithers, L.G., Brinkman, S. and Lynch, J.W. (2016). How Well can Poor Child Development be Predicted from Early Life Characteristics? : A Whole-of-population Data Linkage Study. *Early Childhood Research Quarterly* 35, pp.19-30.

Currie, J. et Goodman, J. (2010). Parental Socioeconomic Status, Child Health, and Human Capital. *International Encyclopedia of Education* 2, 253-259. Copy at http://j.mp/1esWJxz

Currie, J. et Moretti, E. (2003). Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings. *Quarterly Journal of Economics* 118(4), 1495-1532.

Currie, J. et Stabile, M. (2003). Socioeconomic Status and Child Health: Why Is the Relationship Stronger for Older Children? *American Economic Review* 93(5), 1813-1823.

Contoyannis. P. et Li. J. (2011). The Evolution of Health Outcomes from Childhood to Adolescence. *Journal of Health Economics* 30(1), 11-32.

Committee of Environmental Health (2004). Ambient Air Pollution: Health Hazards to Children. *American Academy of Pediatrics* 114, 1699-1707.

Cutler, D. et Lleras-Muney, A. (2006). Education and Health: Evaluating theories and Evidence. *NBER working papers.* No.12352.

Feldman, K. A. (1989). The Association between Student Rating of Specific Dimensions and Student Achievement: Refining and Extending the Synthesis of Data from Multisection Validity Study. *Journal of Research in Higher Education* 30 (6), 583 - 645.

Feinstein, J.S. (1993). The Relationship between Socioeconomic Status and Health: A Review of the Literature. *The Milbank Quarterly* 71(2), 279-322.

Goeman, J., Meijer, R., et Chaturvedi, N. (2016). L1 and L2 Penalized Regression Models.

Grossman, M. (1972). On the Concept of Health Capital and the Demand for Health. *Journal of Political Economy* 80(2), 223-255.

Grossman, M. Chou, S-Y, Liu, J-T. et Joyce, T. (2010). Parental Education and Child Health: Evidence from a Natural Experiment in Taiwan. American Economic Journal of Applied Economics 2(1), 33-61.

Heckman, J. J. (2006). Skill Formation and the Economics of Investing in Disadvantaged Children. *Science* 312(5782), 1900-1902.

James, G., Witten, D., Hastie, T. et Tibshirani, R. (2014). *An Introduction to Statistical Learning,* Collection: Springer Texts in Statistics. New-York: Éditions Springer.

Kitagawa, E.M. et Hauser, P.M. (1973). *Differential Mortality in the United States: A Study in Socioeconomic Epidemiology*. Cambridge. MA: Harvard UP.

Larson, C.P. (2007). Poverty during Pregnancy: Its Effects on Child Health Outcomes. *Paediatrics and Child Health* 12(8), 673–677.

Lawson, J., Janssen. I., Bruner. M., Hossain. A., et Pickett. W. (2014). Asthma Incidence and Risk Factors in a National Longitudinal Sample of Adolescent Canadians: A Prospective Cohort Study. *BMC Pulmonary Medicine* 14, 51.

Marshall, A. (1920). *Principles of Economics* (8e éd.). Londres. Macmillan and Co.

Mullainathan, S., et Spiess, J. (2017). Machine Learning: an Applied Econometric Approach. *Journal of Economic Perspectives* 31(2), 87-106.

Palloni, A., Milesi, C., Turner, A. et White, R.G. (2009). Early Childhood Health. Reproduction of Economic Inequalities and the Persistence of Health and Mortality Differentials. *Social Science and Medicine Review* 68, 1574 – 1582.

Samet, J. et Maynard, R. (2005). Susceptibility of Children to Air Pollution. Dans *Effects of Air Pollution on Children's Health and Development,* p.11-14. Copenhague: World Health Organization.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science* 25(3), 289-310.

Salam, M. T., Millstein, J., Li, Y.-F., Lurmann, F. W., Margolis, H. G., & Gilliland, F. D. (2005). Birth Outcomes and Prenatal Exposure to Ozone, Carbon Monoxide, and Particulate Matter: Results from the Children's Health Study. *Environmental Health Perspectives*, 113(11), 1638–1644. http://doi.org/10.1289/ehp.8111

Scheffer, J. (2002). Dealing with missing data, *Research Letters in the Information and Mathematical Sciences* 3, 153-160.

Smith, J. (1998). Socioeconomic Status and Health. *The American Economic Review* 88(2), 192-196.

Smith, J.P. et Kington, R.S. (1997). Race, Socioeconomic Status, and Health in Late Life. *National Research Council (US) Committee on Population*; avec Martin, L.G. et Soldo, B.J. (rédacteurs). Racial and Ethnic Differences in the Health of Older Americans. Washington (DC): National Academies Press (US); 1997, 4. http://www.ncbi.nlm.nih.gov/books/NBK109835/

Statistique Canada (2009). *Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ).* Récupéré de http://www23.statcan.gc.ca/imdb/p2SV_f.pl?Function=getSurvey\&SDDS=4450.

Statistique Canada (1995 à 2009). *Guide du chercheur de l'enquête longitudinale nationale sur les enfants et les jeunes.* Récupéré de http://www23.statcan.gc.ca/imdb-bmdi/document/4450_D2_T9_V4-fra.pdf.

Statistique Canada. Tableau 102-0122 - Espérance de vie en fonction de la santé. à la naissance et à 65 ans, selon le sexe et le revenu. Canada et provinces. Occasionnel (années). CANSIM (base de données). (site consulté : http://www5.statcan.gc.ca/cansim/a26?lang=fra&retrLang=fra&id=1020122&pattern=espE0rance+de+vie&tabMode=dataTable&srchLan=-1&p1=1&p2=-1)

Statistique Canada (2006). *Fichier de Conversion des codes postaux.* Récupéré de http://geodepot.statcan.ca/2006/Reference/Freepub/92-153-GWF/2007002/using_f.htm

Statistique Canada (2016). *Cote air santé (CAS) et l'indice de qualité de l'air (IQA).* Récupéré de https://ec.gc.ca/cas-aqhi/default.asp?lang=Fr&n=22BA50A8-1

Statistique Canada (2013). *Réseau national de surveillance de la pollution atmosphérique.* Récupéré de https://ec.gc.ca/rnspa-naps/Default.asp?lang=Fr&n=5C0D33CF-1

Stieb, D., Burnette, T., Smith-Doiron, M., Brion O., Shin, H.H. et Economou, V. A. (2008). New Multipollutant, No-Threshold Air Quality Health Index Based on Short-Term Associations Observed in Daily Time-Series Analyses, *Journal of the Air & Waste Management Association* 58(5), 435-450.

Tape, T. T. MD (2017). *Interpreting Diagnistic Tests.* Notes de cours: University of Nebraska Medical Center, Récupéré de http://gim.unmc.edu/dxtests/

Wooldridge, J.M. (2012). *Introductory Econometrics; a Modern Approach.* (5e éd.) South-Western CENGAGE Learning, 879 pages.

# ANNEXE

TABLEAU A1. Résultat des modèles Logit et Probit sans variables environnementales

| | Logit | | | | Probit | | | |
|---|---|---|---|---|---|---|---|---|
| | Cycle 2 | Cycle 3 | Cycle 4 | Cycle 5 | Cycle 2 | Cycle 3 | Cycle 4 | Cycle 5 |
| Année naiss. | 0,294 | 0,205 | 0,109 | -0,060 | 0,148 | 0,112 | 0,066 | -0,042 |
| | (0,165) | (0,175) | (0,187) | (0,192) | (0,088) | (0,094) | (0,101) | (0,104) |
| Mois de naiss. | 0,048 | -0,008 | -0,029 | 0,027 | 0,025 | -0,007 | -0,016 | 0,016 |
| | (0,027) | (0,029) | (0,030) | (0,033) | (0,014) | (0,015) | (0,016) | (0,018) |
| Dépression parentale | 0,594 | 0,356 | 0,816 | 0,800 | 0,338 | 0,194 | 0,474 | 0,435 |
| | (0,221) | (0,262) | (0,266) | (0,297) | (0,124) | (0,146) | (0,152) | (0,169) |
| Santé naissance | 0,823 | 0,575 | 0,734 | 0,441 | 0,463 | 0,307 | 0,423 | 0,219 |
| | (0,249) | (0,281) | (0,297) | (0,339) | (0,138) | (0,157) | (0,169) | (0,188) |
| Naissance prématurée | 0,073 | 0,400 | -0,303 | 0,268 | 0,051 | 0,201 | -0,168 | 0,158 |
| | (0,189) | (0,193) | (0,227) | (0,228) | (0,101) | (0,105) | (0,122) | (0,123) |
| Mère tabac grossesse | -0,263 | 0,770 | -0,048 | 0,355 | -0,152 | 0,415 | -0,039 | 0,171 |
| | (0,280) | (0,328) | (0,347) | (0,396) | (0,153) | (0,175) | (0,188) | (0,213) |
| Drogues pharma. | -0,355 | 0,238 | -0,007 | -0,366 | -0,185 | 0,117 | 0,016 | -0,197 |
| | (0,201) | (0,200) | (0,225) | (0,252) | (0,106) | (0,108) | (0,122) | (0,134) |
| Mère cigarette | 0,305 | -0,462 | 0,007 | -0,499 | 0,167 | -0,265 | 0,015 | -0,253 |
| | (0,263) | (0,325) | (0,322) | (0,375) | (0,143) | (0,172) | (0,174) | (0,200) |
| Asthme | 1,343 | 0,903 | 1,524 | -0,250 | 0,754 | 0,507 | 0,878 | -0,125 |
| | (0,300) | (0,342) | (0,338) | (0,564) | (0,174) | (0,197) | (0,199) | (0,305) |
| Mère mariée | -0,092 | -0,410 | -0,564 | -0,099 | -0,062 | -0,218 | -0,317 | -0,051 |
| | (0,257) | (0,267) | (0,281) | (0,315) | (0,139) | (0,146) | (0,156) | (0,173) |
| Revenu imp. | -0,108 | 0,132 | 0,048 | -0,242 | -0,045 | 0,058 | 0,025 | -0,143 |
| | (0,259) | (0,276) | (0,289) | (0,316) | (0,137) | (0,148) | (0,156) | (0,168) |
| Revenu | 0,162 | -0,166 | -0,243 | -0,371 | 0,091 | -0,089 | -0,131 | -0,216 |
| | (0,125) | (0,130) | (0,144) | (0,153) | (0,066) | (0,070) | (0,077) | (0,083) |
| Éducation mère | -0,197 | 0,089 | -0,048 | -0,388 | -0,108 | 0,043 | -0,043 | -0,221 |
| | (0,200) | (0,210) | (0,226) | (0,240) | (0,107) | (0,114) | (0,123) | (0,131) |
| Éducation conjoint | -0,129 | -0,541 | -0,379 | 0,111 | -0,092 | -0,293 | -0,217 | 0,067 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | (0,208) | (0,210) | (0,225) | (0,252) | (0,110) | (0,112) | (0,122) | (0,135) |
| Deux parents bio. | 0,548 | 14,496 | 1,028 | -1,361 | 0,375 | 4,654 | 0,605 | -0,806 |
| | (1,156) | (561,364) | (1,334) | (1,576) | (0,647) | (88,389) | (0,787) | (0,953) |
| Grosseur famille | -0,143 | 0,005 | -0,381 | 0,663 | -0,093 | 0,020 | -0,242 | 0,370 |
| | (0,697) | (0,640) | (0,772) | (0,817) | (0,378) | (0,351) | (0,414) | (0,447) |
| Mère immigrante | -0,098 | -0,093 | 0,064 | -0,907 | -0,067 | -0,078 | 0,035 | -0,475 |
| | (0,361) | (0,401) | (0,451) | (0,624) | (0,194) | (0,216) | (0,244) | (0,310) |
| Conjoint immigrant | 0,700 | 0,367 | -0,082 | 0,023 | 0,376 | 0,201 | -0,044 | 0,066 |
| | (0,327) | (0,351) | (0,412) | (0,501) | (0,179) | (0,192) | (0,221) | (0,258) |
| Dépression parent. | -0,833 | 0,163 | -0,232 | -0,310 | -0,399 | 0,070 | -0,123 | -0,178 |
| | (0,335) | (0,252) | (0,311) | (0,322) | (0,159) | (0,134) | (0,161) | (0,167) |
| Petit poids (naiss.) | 0,346 | 0,938 | 0,583 | 0,482 | 0,174 | 0,534 | 0,364 | 0,235 |
| | (0,431) | (0,412) | (0,449) | (0,516) | (0,237) | (0,233) | (0,252) | (0,281) |
| Santé Mère | 0,958 | 0,693 | 0,520 | 0,478 | 0,522 | 0,375 | 0,294 | 0,272 |
| | (0,191) | (0,206) | (0,232) | (0,258) | (0,105) | (0,114) | (0,130) | (0,144) |
| Naiss. prématurée | -0,451 | -0,421 | -0,014 | -0,697 | -0,224 | -0,228 | -0,029 | -0,379 |
| | (0,346) | (0,360) | (0,355) | (0,452) | (0,185) | (0,197) | (0,197) | (0,241) |
| Taille pop. | 0,024 | 0,063 | 0,030 | -0,095 | 0,016 | 0,033 | 0,016 | -0,052 |
| | (0,080) | (0,086) | (0,096) | (0,106) | (0,043) | (0,046) | (0,052) | (0,056) |
| Est (loc.) | -0,007 | -1,587 | -0,169 | 0,015 | 0,023 | -0,836 | -0,095 | 0,025 |
| | (0,464) | (0,377) | (0,479) | (0,537) | (0,245) | (0,211) | (0,260) | (0,292) |
| Québec (loc.) | 0,438 | -1,229 | -0,378 | 0,190 | 0,238 | -0,664 | -0,228 | 0,123 |
| | (0,452) | (0,374) | (0,495) | (0,538) | (0,241) | (0,212) | (0,269) | (0,294) |
| Ontario (loc.) | 0,214 | -1,313 | 0,046 | 0,300 | 0,150 | -0,720 | 0,025 | 0,197 |
| | (0,443) | (0,360) | (0,471) | (0,526) | (0,236) | (0,204) | (0,256) | (0,287) |
| Prairies (loc.) | -0,041 | -1,494 | -0,008 | -0,231 | 0,038 | -0,822 | -0,016 | -0,095 |
| | (0,459) | (0,381) | (0,483) | (0,552) | (0,243) | (0,215) | (0,263) | (0,298) |
| Alberta (loc.) | 0,196 | -1,128 | -0,555 | -0,096 | 0,142 | -0,614 | -0,361 | -0,048 |
| | (0,498) | (0,428) | (0,575) | (0,603) | (0,264) | (0,240) | (0,308) | (0,326) |
| C.B. (loc.) | 0,719 | -1,388 | -0,848 | -0,663 | 0,422 | -0,729 | -0,500 | -0,325 |
| | (0,495) | (0,449) | (0,625) | (0,688) | (0,267) | (0,248) | (0,329) | (0,363) |
| | 2141 | 1855 | 1535 | 1352 | 2141 | 1855 | 1535 | 1352 |

Variables non-significatives, non présentées dans le tableau; Nombre de grossesses de la mère, l'âge de la mère à la naissance, la Fratrie, l'utilisation des drogues prescrites de la mère, le tabac pour le conjoint, le type de logement familial, que la mère buvait durant la grossesse, l'absence du conjoint et le sexe de l'enfant

TABLEAU A2. Résultats des régressions Logit et Probit, avec les variables environnementales

| | Logit | | | | Probit | | | |
|---|---|---|---|---|---|---|---|---|
| | Cycle 2 | Cycle 3 | Cycle 4 | Cycle 5 | Cycle 2 | Cycle 3 | Cycle 4 | Cycle 5 |
| Année naiss. | 0,260 | 0,179 | 0,142 | -0,079 | 0,134 | 0,091 | 0,071 | -0,058 |
| | (0,171) | (0,183) | (0,198) | (0,204) | (0,091) | (0,098) | (0,107) | (0,112) |
| Mois naiss. | 0,046 | -0,003 | -0,023 | 0,051 | 0,025 | -0,005 | -0,014 | 0,027 |
| | (0,028) | (0,030) | (0,032) | (0,035) | (0,015) | (0,016) | (0,017) | (0,019) |
| Fratrie | 0,317 | 0,118 | -0,137 | -0,547 | 0,182 | 0,064 | -0,059 | -0,294 |
| | (0,221) | (0,213) | (0,240) | (0,276) | (0,122) | (0,117) | (0,132) | (0,149) |
| Âge mère naiss. | -0,052 | 0,008 | 0,005 | 0,024 | -0,030 | 0,005 | 0,004 | 0,011 |
| | (0,023) | (0,023) | (0,026) | (0,028) | (0,013) | (0,013) | (0,014) | (0,015) |
| Dépression parent. | 0,571 | 0,407 | 0,866 | 0,912 | 0,311 | 0,208 | 0,498 | 0,494 |
| | (0,232) | (0,270) | (0,277) | (0,319) | (0,129) | (0,151) | (0,157) | (0,180) |
| Santé à la naissance | 0,941 | 0,713 | 0,716 | 0,359 | 0,517 | 0,363 | 0,413 | 0,201 |
| | (0,263) | (0,294) | (0,321) | (0,358) | (0,145) | (0,164) | (0,180) | (0,197) |
| Prématuré | 0,175 | 0,386 | -0,356 | 0,287 | 0,109 | 0,194 | -0,205 | 0,186 |
| | (0,196) | (0,201) | (0,239) | (0,243) | (0,105) | (0,109) | (0,128) | (0,132) |
| Mère tabac | -0,178 | 0,859 | -0,100 | 0,310 | -0,090 | 0,443 | -0,078 | 0,170 |
| | (0,293) | (0,343) | (0,357) | (0,412) | (0,159) | (0,184) | (0,195) | (0,225) |
| Drogues Presc. | 0,251 | 0,053 | -0,123 | -0,356 | 0,161 | 0,035 | -0,079 | -0,191 |
| | (0,203) | (0,209) | (0,235) | (0,260) | (0,109) | (0,113) | (0,128) | (0,141) |
| Drogues Pharma. | -0,340 | 0,330 | 0,004 | -0,473 | -0,185 | 0,146 | 0,030 | -0,269 |
| | (0,208) | (0,211) | (0,238) | (0,269) | (0,110) | (0,114) | (0,128) | (0,145) |
| Mère cigarette | 0,279 | -0,592 | 0,061 | -0,470 | 0,135 | -0,310 | 0,047 | -0,258 |
| | (0,275) | (0,338) | (0,331) | (0,390) | (0,149) | (0,179) | (0,180) | (0,212) |
| Grossesses | -0,082 | 0,119 | 0,108 | 0,121 | -0,048 | 0,063 | 0,048 | 0,065 |
| | (0,096) | (0,106) | (0,106) | (0,127) | (0,053) | (0,058) | (0,059) | (0,070) |
| Asthme | 1,378 | 0,806 | 1,484 | -0,356 | 0,751 | 0,459 | 0,866 | -0,230 |
| | (0,316) | (0,357) | (0,356) | (0,590) | (0,181) | (0,205) | (0,207) | (0,324) |
| Mère mariée | -0,205 | -0,420 | -0,593 | -0,116 | -0,093 | -0,205 | -0,335 | -0,065 |
| | (0,271) | (0,280) | (0,299) | (0,348) | (0,147) | (0,153) | (0,165) | (0,191) |
| Revenu | 0,211 | -0,159 | -0,238 | -0,433 | 0,113 | -0,092 | -0,132 | -0,247 |
| | (0,129) | (0,133) | (0,149) | (0,162) | (0,069) | (0,072) | (0,081) | (0,089) |
| Éduc. mère | -0,312 | 0,082 | -0,157 | -0,505 | -0,168 | 0,048 | -0,127 | -0,282 |
| | (0,210) | (0,218) | (0,238) | (0,261) | (0,112) | (0,118) | (0,129) | (0,143) |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Éducation conjoint | -0,188 | -0,531 | -0,479 | 0,140 | -0,111 | -0,281 | -0,264 | 0,083 |
| | (0,216) | (0,216) | (0,236) | (0,264) | (0,115) | (0,116) | (0,128) | (0,143) |
| Grosseur famille | -0,082 | -0,044 | -0,163 | 0,793 | -0,100 | -0,011 | -0,151 | 0,466 |
| | (0,718) | (0,648) | (0,788) | (0,899) | (0,392) | (0,361) | (0,429) | (0,489) |
| Dépression parentale | -0,857 | 0,266 | -0,264 | -0,361 | -0,425 | 0,140 | -0,121 | -0,217 |
| | (0,345) | (0,264) | (0,323) | (0,334) | (0,166) | (0,139) | (0,168) | (0,177) |
| Naiss. Petit poids | 0,123 | 0,959 | 0,611 | 0,539 | 0,077 | 0,531 | 0,373 | 0,291 |
| | (0,456) | (0,433) | (0,478) | (0,542) | (0,249) | (0,244) | (0,268) | (0,295) |
| Alcool mère grossesse | -0,008 | 0,376 | 0,358 | -0,201 | 0,000 | 0,207 | 0,193 | -0,108 |
| | (0,242) | (0,258) | (0,278) | (0,310) | (0,131) | (0,140) | (0,153) | (0,168) |
| Santé Mère | 0,953 | 0,642 | 0,472 | 0,494 | 0,518 | 0,351 | 0,251 | 0,293 |
| | (0,199) | (0,218) | (0,246) | (0,281) | (0,109) | (0,120) | (0,137) | (0,155) |
| Prématuré | -0,472 | -0,470 | 0,125 | -0,696 | -0,231 | -0,250 | 0,033 | -0,387 |
| | (0,364) | (0,378) | (0,375) | (0,477) | (0,194) | (0,207) | (0,209) | (0,255) |
| Est (loc.) | 0,378 | -1,428 | -0,200 | -0,272 | 0,226 | -0,712 | -0,129 | -0,118 |
| | (0,622) | (0,540) | (0,667) | (0,718) | (0,323) | (0,291) | (0,355) | (0,392) |
| Québec (loc.) | 0,108 | -1,649 | -0,861 | 0,703 | 0,085 | -0,827 | -0,467 | 0,440 |
| | (0,645) | (0,609) | (0,722) | (0,755) | (0,349) | (0,339) | (0,395) | (0,416) |
| Ontario (loc.) | 0,297 | -1,415 | -0,643 | 0,166 | 0,160 | -0,736 | -0,321 | 0,228 |
| | (0,819) | (0,797) | (0,969) | (0,968) | (0,437) | (0,438) | (0,521) | (0,529) |
| Prairies (loc.) | 0,195 | -1,633 | 1,240 | 0,577 | 0,140 | -0,808 | 0,679 | 0,445 |
| | (1,513) | (1,542) | (1,825) | (1,813) | (0,815) | (0,844) | (0,990) | (0,992) |
| Moyenne ($SO_2$) | -0,032 | -0,007 | -0,030 | 0,029 | -0,014 | -0,001 | -0,021 | 0,014 |
| | (0,078) | (0,079) | (0,094) | (0,099) | (0,041) | (0,042) | (0,050) | (0,054) |
| Écart-Type ($SO_2$) | 0,122 | 0,067 | -0,153 | -0,233 | 0,060 | 0,031 | -0,089 | -0,132 |
| | (0,234) | (0,227) | (0,260) | (0,321) | (0,123) | (0,122) | (0,140) | (0,171) |
| Captation ($NO_2$) | -0,402 | 0,323 | 0,091 | 0,106 | -0,225 | 0,191 | 0,049 | 0,066 |
| | (0,170) | (0,204) | (0,188) | (0,196) | (0,089) | (0,109) | (0,100) | (0,104) |
| Moyenne ($NO_2$) | -0,120 | 0,075 | 0,328 | 0,294 | -0,051 | 0,049 | 0,188 | 0,181 |
| | (0,144) | (0,136) | (0,149) | (0,163) | (0,075) | (0,073) | (0,081) | (0,090) |
| Écart-Type ($NO_2$) | 0,246 | -0,015 | -0,203 | -0,568 | 0,120 | -0,038 | -0,106 | -0,325 |
| | (0,185) | (0,199) | (0,231) | (0,261) | (0,098) | (0,106) | (0,121) | (0,139) |
| Moyenne ($O_3$) | -0,096 | 0,132 | 0,109 | 0,135 | -0,053 | 0,073 | 0,059 | 0,070 |
| | (0,093) | (0,093) | (0,103) | (0,109) | (0,049) | (0,050) | (0,056) | (0,060) |
| Écart-Type ($O_3$) | 0,061 | -0,316 | -0,205 | -0,347 | 0,039 | -0,178 | -0,111 | -0,182 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | (0,175) | (0,190) | (0,217) | (0,219) | (0,093) | (0,102) | (0,115) | (0,118) |
| Captation (NO) | -1,532 | -0,727 | -0,473 | -0,529 | -0,755 | -0,382 | -0,231 | -0,225 |
| | (1,032) | (1,043) | (1,054) | (1,153) | (0,536) | (0,546) | (0,571) | (0,630) |
| Moyenne (NO) | -0,174 | -0,065 | 0,272 | 0,278 | -0,075 | -0,027 | 0,170 | 0,179 |
| | (0,191) | (0,189) | (0,213) | (0,235) | (0,100) | (0,102) | (0,115) | (0,129) |
| Écart Type (NO) | 0,589 | 0,061 | 0,107 | -0,700 | 0,300 | 0,011 | 0,044 | -0,399 |
| | (0,244) | (0,249) | (0,279) | (0,335) | (0,128) | (0,132) | (0,148) | (0,178) |
| Captation ($NO_x$) | 1,958 | 0,395 | 0,392 | 0,427 | 0,992 | 0,186 | 0,188 | 0,162 |
| | (0,997) | (0,993) | (1,025) | (1,120) | (0,519) | (0,520) | (0,555) | (0,612) |
| Moyenne ($NO_x$) | 0,096 | 0,002 | -0,305 | -0,262 | 0,034 | -0,006 | -0,184 | -0,170 |
| | (0,182) | (0,175) | (0,197) | (0,215) | (0,095) | (0,094) | (0,106) | (0,118) |
| Écart-Type ($NO_x$) | -0,497 | 0,043 | -0,089 | 0,587 | -0,252 | 0,044 | -0,035 | 0,335 |
| | (0,253) | (0,258) | (0,301) | (0,336) | (0,133) | (0,136) | (0,159) | (0,179) |
| Agricole | 0,652 | -0,223 | 2,018 | 0,853 | 0,307 | -0,083 | 1,124 | 0,501 |
| | (0,745) | (0,964) | (0,903) | (0,870) | (0,388) | (0,470) | (0,470) | (0,470) |
| Commercial | 1,199 | 1,929 | 2,585 | 1,327 | 0,654 | 0,981 | 1,420 | 0,745 |
| | (0,700) | (0,744) | (0,864) | (0,901) | (0,372) | (0,396) | (0,463) | (0,493) |
| Forestier | 0,897 | 1,380 | 1,952 | 0,705 | 0,457 | 0,707 | 1,011 | 0,376 |
| | (0,598) | (0,651) | (0,747) | (0,755) | (0,321) | (0,350) | (0,401) | (0,410) |
| Industriel | 0,013 | 1,344 | 1,237 | 2,460 | -0,052 | 0,773 | 0,720 | 1,401 |
| | (1,048) | (0,965) | (1,239) | (1,146) | (0,546) | (0,517) | (0,660) | (0,637) |
| Résidentiel | 0,642 | 1,915 | 2,566 | 2,506 | 0,342 | 1,006 | 1,416 | 1,400 |
| | (0,667) | (0,709) | (0,818) | (0,856) | (0,352) | (0,377) | (0,436) | (0,466) |
| Montréal | 0,201 | 1,577 | 1,642 | 0,661 | 0,199 | 0,848 | 0,788 | 0,555 |
| | (0,992) | (0,891) | (1,264) | (1,405) | (0,512) | (0,492) | (0,671) | (0,712) |
| Distance | -0,002 | 0,001 | -0,001 | -0,001 | -0,001 | 0,000 | 0,000 | 0,000 |
| | (0,001) | (0,001) | (0,002) | (0,002) | (0,001) | (0,001) | (0,001) | (0,001) |
| N | 2141 | 1855 | 1535 | 1352 | 2141 | 1855 | 1535 | 1352 |

Variables inclues dans les régressions mais non-démontrées et non-significatives; Élévation, Sudburry, Alberta, Bc, conjoint cigarette, type de logement, revenu imputé, sexe de l'enfant, PCM est immigrant, Conjoint est immigrant, Complete 4, Complete $O_3$, Calgary, Edmonton, Hamilton, Vancouver, Longitude et latitude, taille de population, absence du conjoint Deux parents biologiques et la constante.

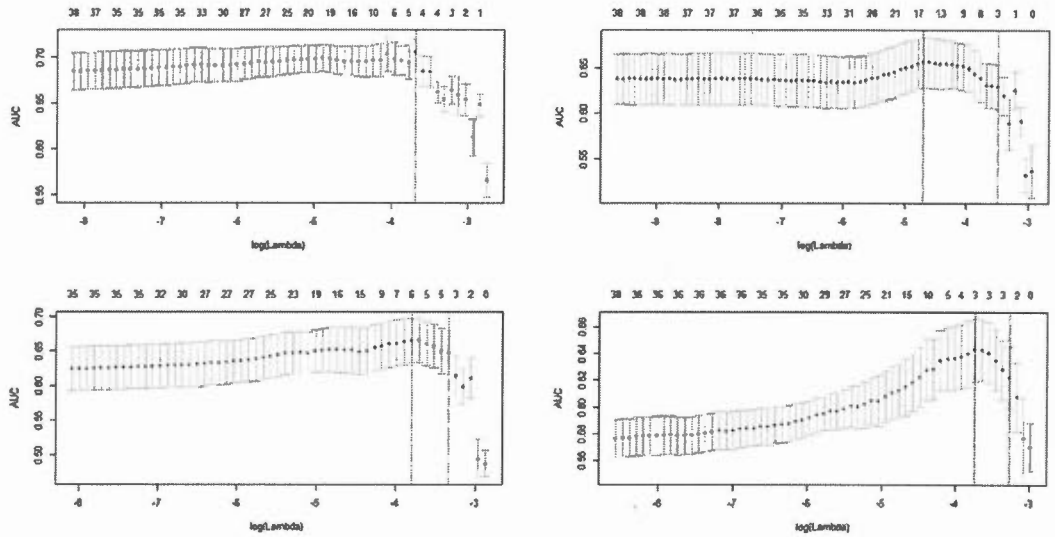


FIGURE A1. Choix de λ optimal pour le *lasso* en validation croisée sans les variables environnementales

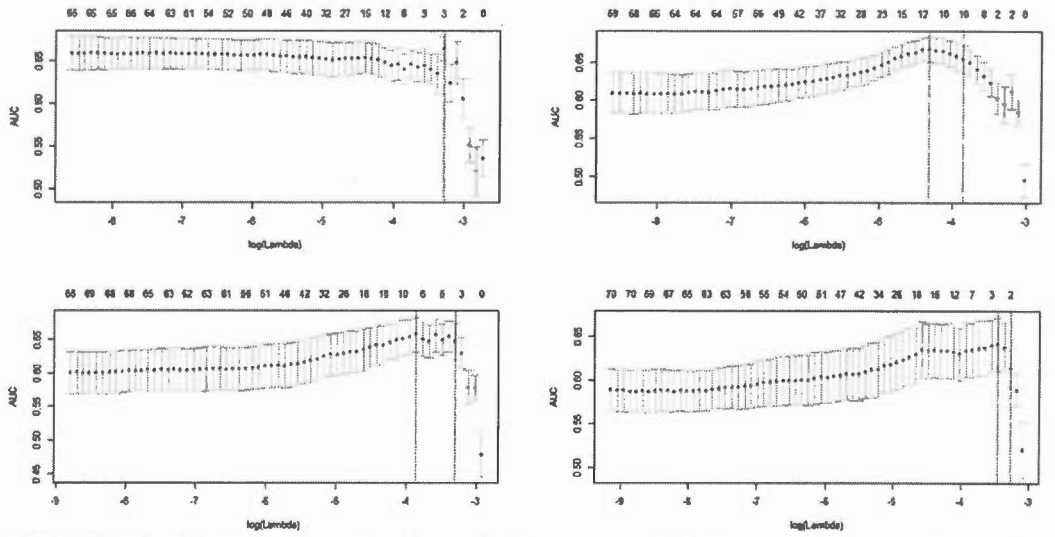


FIGURE A2. Choix de λ optimal pour le *lasso* en validation croisée avec les variables environnementales

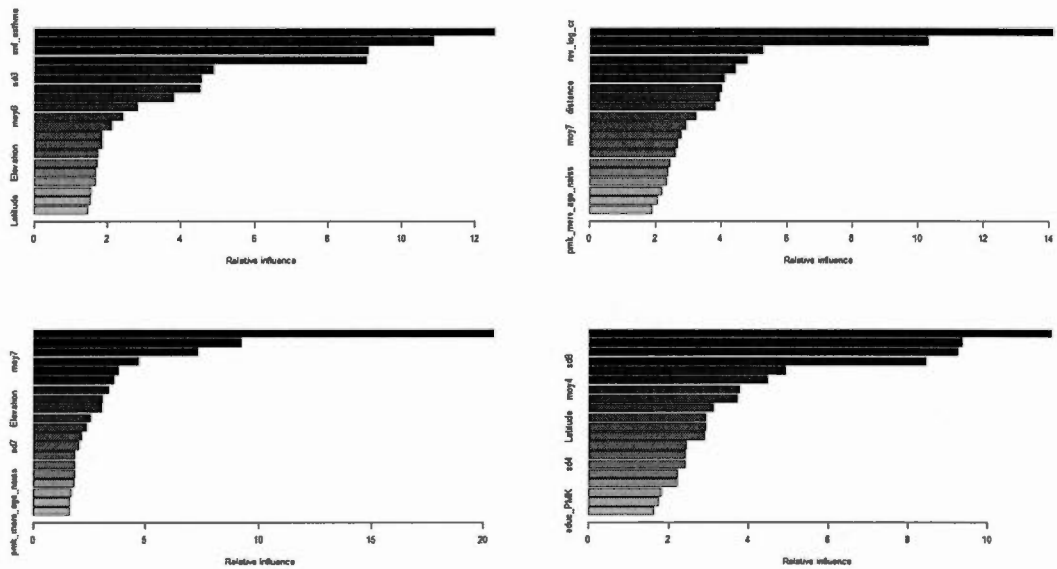



FIGURE A.3 : Influence relative des modèles Boosting avec variables environnementales

TABLEAU A.3 : Variables principales de la méthode *boosting* par ordre d'influence relative avec les variables environnementales

| Cycle 2 | | Cycle 3 | | Cycle 4 | | Cycle 5 | |
|---|---|---|---|---|---|---|---|
| Asthme | 12,5 | Santé Mère | 14,1 | Asthme | 20,4 | Revenu | 11,6 |
| Santé Naissance | 10,9 | Revenu | 10,3 | Santé Mère | 9,2 | Santé Mère | 9,4 |
| Santé Mère | 9,1 | Mère tabac | 5,3 | Revenu | 7,3 | Dépression parent. | 9,3 |
| Écart-type ($O_3$) | 9,0 | Mère mariée | 4,8 | Moyenne ($O_3$) | 4,7 | Écart Type (NO) | 8,4 |
| Moyenne ($SO_2$) | 4,9 | Écart-T. ($NO_2$) | 4,4 | Dépression parent | 3,8 | Écart-Type ($NO_2$) | 4,9 |
| Écart-Type ($NO_2$) | 4,6 | Nbre grossesses | 4,1 | Éduc. conjoint | 3,6 | distance | 4,5 |
| Grosseur famille | 4,5 | Moyenne ($NO_2$) | 4,0 | Mère mariée | 3,3 | Moyenne ($SO_2$) | 3,8 |
| distance | 3,8 | distance | 3,9 | Moyenne ($SO_2$) | 3,1 | Captation ($O_3$) | 3,7 |
| Grossesse | 2,8 | Mois de naiss. | 3,8 | Élevation | 3,1 | Complete4 | 3,1 |
| Moyenne (NO) | 2,4 | Moyenne ($SO_2$) | 3,2 | Latitude | 2,5 | Moyenne (NOx) | 2,9 |
| Captation ($O_3$) | 2,1 | Latitude | 2,9 | Santé Naissance | 2,4 | Latitude | 2,9 |
| Forestier | 1,9 | Moyenne ($O_3$) | 2,8 | Écart-Type ($SO_2$) | 2,1 | Québec (loc.) | 2,9 |
| Mère cigarette | 1,8 | Asthme | 2,7 | É.-T. ($O_3$) | 2,0 | Moyenne (NO) | 2,4 |
| Captation ($NO_2$) | 1,7 | Moyenne ($NO_x$) | 2,6 | Captation ($NO_2$) | 1,8 | Moyenne ($NO_2$) | 2,4 |
| Élevation | 1,7 | Moyenne (NO) | 2,4 | Éduc. mère | 1,8 | Écart-Type ($SO_2$) | 2,4 |

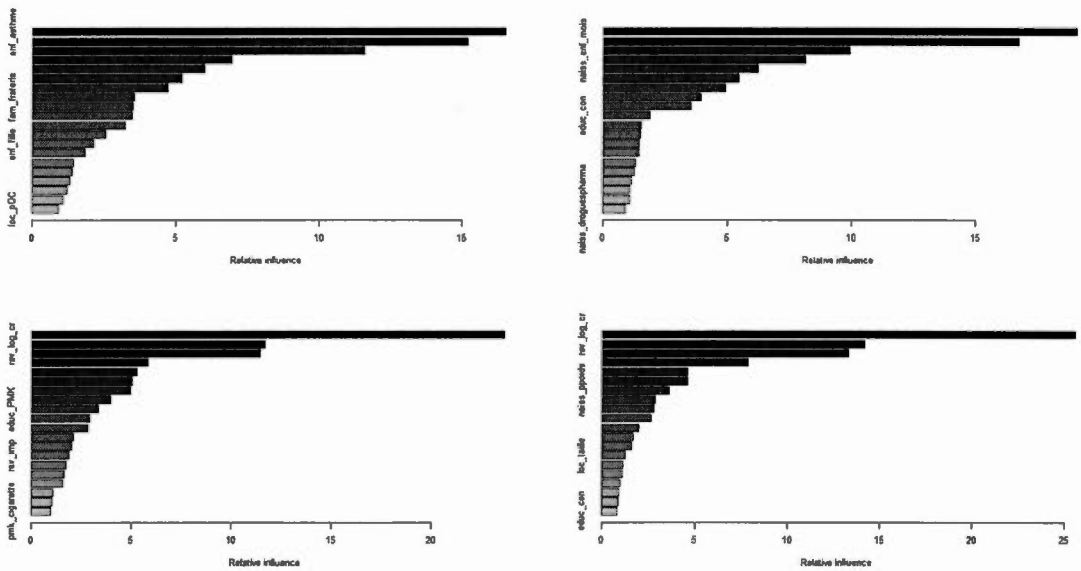



FIGURE A.4 : Influence relative des variables du modèle *boosting*, sans les variables environnementales

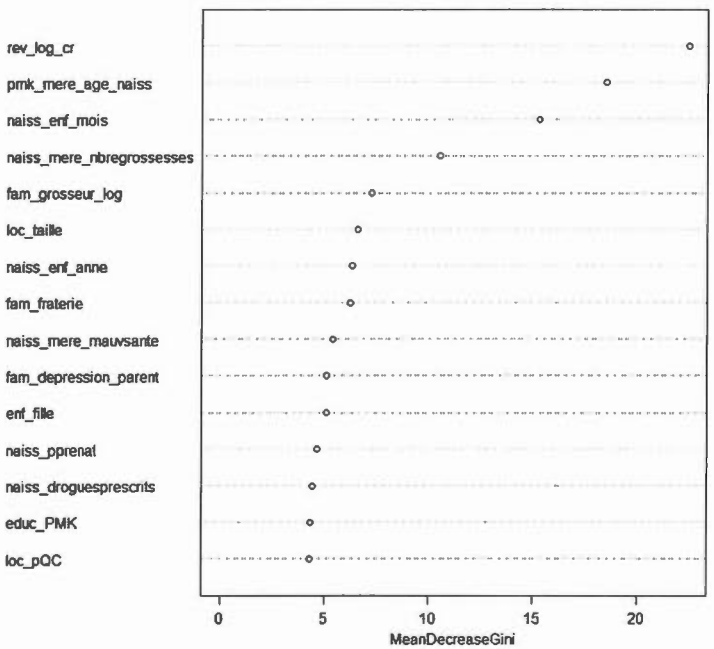



FIGURE A.5 Influence des variables de forêt aléatoire, sans variables environnementales au cycle 5

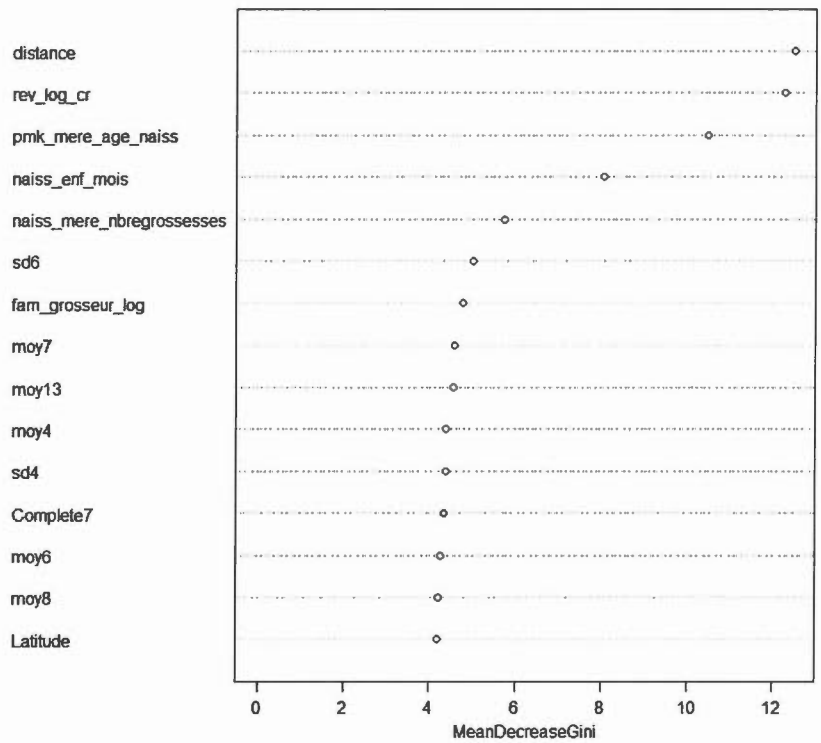



FIGURE A.6 Influence des variables de forêt aléatoire, avec variables environnementales au cycle 5

TABLEAU A.4 Nombre d'arbres de la méthode *boosting*

| | Cycle 2 | Cycle 3 | Cycle 4 | Cycle 5 |
|---|---|---|---|---|
| Avec variables environnementales | 1227 | 885 | 883 | 736 |
| Sans variables environnementales | 1214 | 1120 | 962 | 730 |